Supporting Information related to

DNA methylation profiling reveals a predominant immune component in breast cancers

Sarah Dedeurwaerder, Christine Desmedt, Emilie Calonne, Sandeep K. Singhal, Benjamin Haibe-Kains, Matthieu Defrance, Stefan Michiels, Michael Volkmar, Rachel Deplus, Judith Luciani, Françoise Lallemand, Denis Larsimont, Jérôme Toussaint, Sandy Haussy, Françoise Rothé, Ghizlane Rouas, Otto Metzger, Samira Majjaj, Kamal Saini, Pascale Putmans, Gérald Hames, Nicolas van Baren, Pierre G. Coulie, Martine Piccart, Christos Sotiriou & François Fuks

Supplemental Materials and Methods

Breast cancer "expression subtype" determination

Two approaches were used to determine "breast cancer expression subtypes". First, on the basis of an IHC determination, basal-like tumours were defined as negative for ER and HER2 receptors and as histological grade 3, HER2 tumours as overexpressing the HER2 receptor, and luminal tumours as ER positive and HER2 negative. This last group was divided into luminal A and B tumours corresponding respectively to histological grade 1 and grade 3 tumours. Secondly, the subtypes were identified on the basis of gene expression by applying the Subtype Classification Model as described in (Desmedt et al., 2008) and (Wirapati et al., 2008). The only difference was in the use of the single probes "205225_at", "216836_s_at" and "208079_s_at" instead of the full ESR1, ERBB2 and AURKA modules, respectively. We chose to use this simplified version of the Subtype Classification Model as this model showed excellent performance when applied to the Affymetrix dataset, while reducing the number of genes in the clustering model (data not shown). We used the 'genefu' R package, available on CRAN (http://cran.r-project.org/web/packages/genefu/).

Culture of breast epithelial and lymphoid cell lines

MCF10A cells were cultured in DMEM/F12 (1:1) medium (Gibco); MCF-7, SKBR3 and MDA-MB-231 were cultured in DMEM medium (Gibco); T47D, ZR-75-1 and MDA-MB-361 were cultured in RPMI medium (Gibco); and BT20 were cultured in MEM medium (Gibco). For all breast epithelial cell lines, media were supplemented with 10% fetal calf serum (Gibco). The lymphoid clones CD4+ R12C9 and CD8+ WEIS3E5 were maintained in Isocove Dubelcco medium supplemented with 10% human serum HS54, L-

Arginine, L-Asparagine, L-glutamine, 2-mercaptoéthanol and methyltryptophane as well as with 10 ng/mL of IL-7 and 50 U/mL of IL-2.

Isolation of *ex vivo* lymphocytes

Blood mononuclear cells from an hemochromatosis patient were isolated with density gradient centrifugation using Lymphoprep (Axis-Shield PoCAS, Oslo, Norway), and extensively washed in cold phosphate-buffered saline containing 2 mM EDTA, to eliminate platelets. CD3+ and CD20+ cells were purified with magnetic microbeads using the CD3 Isolation Kit or CD20 Isolation Kit (Miltenyi Biotec, Bergisch Gladbach, Germany) in an AUTOMACS magnetic sorter (Miltenyi), following the manufacturer's instructions. Cell purities were higher than 99% and 92% for the CD3+ and CD20+ cells, respectively, as determined with standard flow cytometry.

Bisulphite genomic sequencing

Methylation status of four CpG sites - cg07471052, cg11566244, cg22498251 and cg09847584 - located respectively near the transcription start sites of the *CDK3*, *GSTP1*, *TWIST1* and *RIMBP2* genes, was examined by bisulphite genomic sequencing applied to 1 normal (N1) and 3 breast cancer (BC10, BC32 and BC109) samples. Primers were designed manually and sequences are provided in Supplementary Table SV. The PCR amplified fragments were purified by *QIAquick® Gel Extraction kit* (Qiagen), cloned into the pCR®II-TOPO® vector (Invitrogen, Carlsbad, CA, USA), and used to transform competent *Escherichia coli* TOP10 cells. Clones were selected by blue/white colonie screening and amplified. Plasmids were purified with the *Qiagen-MiniPrep kit* (Qiagen). The PCR products were sequenced by Genoscreen (Lille, France) and CpG methylation status were analysed with the BiQ Analyzer software as described in (Bock et al., 2005).

Bisulphite pyrosequencing

750 ng of genomic DNA were bisulphite-converted using the EZ DNA Methylation™ kit (Zymo Research) as for DNA methylation profiling. One third of the converted DNA was used as template for each subsequent PCR. To ensure sufficient amount of PCR product for sequencing we performed nested PCRs. PCR primers for pre-amplification (EF, ER primers) were deduced manually or with the help of "BiSearch Primer Design and Search Tool" (http://bisearch.enzim.hu) and checked for tendency to form oligomers, hairpin loops etc. using the Generunner software (version 3.05, Hastings Software Inc.). Primers for nested amplification and sequencing were deduced manually or using PyroMark® Assay Design 2.0 software (Qiagen).

Pre-amplification PCRs were conducted with 3mM $MgCl_2$, 1mM of each dNTP, 12% (v/v) DMSO, 500nM of each primer (EF+ER primers, see Supplementary Table SXXX) and optionally 500mM Betaine in heated-lid thermocyclers under the following conditions: 95°C 3:00; 25 cycles of [94°C 0:30; 51°C 0:40; 72°C 1:30]; 72°C 5:00. Nested amplifications (F, RBio primers) were performed with the HotStarTaq PCR kit (Qiagen) using 2% (v/v) of the pre-amplification PCR as template under the following conditions: 95°C 15:00; 45 cycles of [94°C 0:30; 55°C 0:30; 72°C 0:30]; 72°C 10:00. Amplification success was assessed with agarose gel electrophoresis and pyrosequencing of the PCR products (S primers) was performed with the Pyromark™ Q24 system (Qiagen).

Histopathologic analysis of the lymphocyte infiltration

Histopathologic analysis of tumours in order to evaluate both stromal and intratumoral lymphocyte infiltration was performed on hematoxylin and eosin-stained sections, as previously described (Denkert et al., 2010).

Unsupervised clustering

In a first step, as a completely unsupervised approach, hierarchical clustering was performed on all 123 breast tissues of the main set (119 IDCs and 4 normal breast tissues) on the basis of the 10% most variant CpGs between all samples (see Fig S2). This has been done also for all samples of the validation set (see Fig S15). In a second step, hierarchical clustering was performed only on the 119 IDCs of the main set on the basis of a reduced list of CpGs differentially methylated between IDC and normal tissues identified in Table SIII. Among the 6,309 CpGs identified as being differentially methylated between IDC and normal samples, we chose to work with those showing a 20% methylation difference in at least 30% of the IDCs as compared to the normal breast samples (see Table SVII). This ensured selection of a reasonable number of CpGs (2,985) having potentially informative variance in our dataset and yielded clusters showing good stability. Complete linkage and distance correlations were used for clustering arrays and CpGs. The stability of the clustering was estimated with the 'pvclust' R package (Suzuki and Shimodaira, 2006), available on CRAN (http://cran.r-project.org/web/packages/pvclust/). We measured the uncertainty in hierarchical clustering by bootstrap stability probabilities ranging from 0 to 1, with 0 indicating poor stability and 1 indicating a very high stability. The bootstrap probability value of a cluster is the frequency that it appears in the bootstrap replicates. These stability values quantify how strong a cluster is supported by data. The criteria used to select the 6 methylation clusters reported in this paper were: (i) a stability probability of minimum 0.75, and (ii) a minimum number of samples of 8 (see Fig S5).

Module/signature scores

The calculation of module/signature scores is described in (Desmedt et al., 2008) and (Wirapati et al., 2008). Briefly, a signature score, denoted by R_s, was defined as the weighted combination of all the gene expressions in the corresponding signature:

$$R_s = \frac{\sum_{i \in Q} w_i x_i}{n_Q}$$

where Q is the set of genes in the signature, n_Q is the number of genes in Q, x_i is the expression of gene i, and w_i is either -1 or +1 depending on the sign of the statistic/coefficient published in the original study. For the particular cases of the two divided "ESR1 positive" and "ESR1 negative" modules, w_i is always equal to +1. For DNA methylation data, signature scores were calculated in a manner similar to that of gene expression data with an additional mapping procedure: each CpG probe was mapped to the corresponding gene through Entrez Gene ID. Each signature score was scaled so that quantiles 2.5% and 97.5% equaled -1 and +1, respectively. This scaling was robust to outliers and ensured that the signature score lay approximately within the [-1,+1] interval, allowing comparison of datasets based on different microarray technologies and normalizations.

Annotation of Infinium array in terms of CpG location

Additional annotations of the Infinium array were added to the ones provided by Illumina regarding the location of the CpG (i) *versus* CGI (CpG inside a CGI, CpG island shore, other CpG) and (ii) *versus* promoter classes (High-, Intermediated or Low-CpG-density promoter). They are provided in Table SVI.

CpG location versus CGI

CpGs were classified according to their position relatively to CpG islands (*i.e.* CpG inside a CGI, CpG island shore or other CpG). Two classifications were established, and this in function of the CGI definition used: the UCSC definition (CpG_Island_UCSC classification) or the improved and revisited definition described in (Bock et al., 2007) (CpG_Island_Revisited classification). A CpG was considered as a CpG island shore if it was located inside a 2 kb region around a CGI (as defined in (Irizarry et al., 2009)). A

CpG located neither in a CGI nor in a 2 kb region around a CGI was considered as other CpG. Both classifications are provided in Table SVI; we only used the revisited classification described in (Bock et al., 2007) for all analyses.

CpG location versus promoter classes

Promoters represented on the Infinium array were categorized using their CpG content as defined in (Weber et al., 2007). First, regions from -700 to +500 bp surrounding the transcription start site (TSS) were extracted using the UCSC genome browser data (Rhead et al., 2010). Then, using the DNA sequences corresponding to those promoter fragments, the CpG ratio and the GC content were calculated in sliding windows of 500 bp with 5 bp offsets. Finally, according to the definition provided in (Weber et al., 2007), the promoters were classified as HCPs (High-CpG-density promoters) if a least one 500 bp window contains a CpG ratio > 0.75 and a GC content > 0.55 was found; as LCPs (Low-CpG-density promoters) if no 500 bp window has reached a CpG ratio of 0.48; or as ICPs (Intermediate-CpG-density promoters) otherwise.

Methylation difference criterion

Several indications led us to choose 20% as the methylation difference criterion. First, it seemed that the Infinium assay gave values ranging from 0 to 0.2 for unmethylated CpGs. Second, a recent study has shown that for more than 90% of the loci, the sensitivity of methylation difference detection is 0.2 (Bibikova et al., 2009).

Class comparison analyses in the main set of patients

A two-sided Mann-Whitney test (also called Wilcoxon-Mann-Whitney test) was employed to test the null hypothesis (H_0) assumption of equality of the methylation values in two defined groups of data. The loss of power induced by multiple tests was corrected by the false discovery rate (FDR) approach (Benjamini and Hochberg, 1995).

For normal samples we considered the mean of methylation values, because of the small sample size and the low variance. For tumour samples, because of their higher heterogeneity, we considered the median value, less sensitive to extreme values.

Between IDCs and normal breast tissue samples

A particular CpG was considered hyper- or hypo-methylated in IDCs as compared to normal breast tissue samples according to the following two criteria: 1/ the CpG had to show at least a 20% methylation difference in IDCs as compared to normal breast tissue samples in at least 10% of the IDCs; 2/ to be considered hypermethylated, the CpG had to show at least ten times more hypermethylation events than hypomethylation events in breast cancer. Conversely, to be considered hypomethylated, it had to show at least ten times more hypomethylation events than hypermethylation events in breast cancer.

Between the two main clusters, I and II

CpGs differentially methylated between clusters I and II were determined according to these two criteria: 1/ they had to show a methylation difference of at least 20% between the two groups; 2/ the FDR-corrected Wilcoxon p-value for the concerned CpGs had to be lower than 0.1.

Between each methylation subcluster and normal breast tissue samples

The criteria for determining that a given methylation subcluster showed differential methylation with respect to normal breast tissue samples were: 1/ The CpGs concerned had to show a difference in methylation of at least 20% between the two groups; 2/ the Wilcoxon p-value for the CpGs concerned had to be lower than 0.01. Here, we did not use the FDR criterion as described above, because of the small number of samples composing each group.

Gene Set Enrichment Analysis (GSEA)

GSEA is a powerful analytical method first developed to determine if the members of a given gene set are significantly enriched among the genes most differentially expressed between two sample groups (Mootha et al., 2003). Here we applied this method to both our methylation data and our expression data to assess the possibility that ER biology might be regulated by DNA methylation. For this, we hypothesized that the ESR1 module genes were more highly methylated in cluster I ("ER-negative tumours") than in cluster II ("ER-positive tumours").

For this analysis, the ESR1 module described in (Desmedt et al., 2008) had to be divided into two sub-modules: an ESR1-positive module, containing all ESR1 module genes whose expression correlates positively with ESR1 expression, and an ESR1-negative module containing those whose expression correlates negatively with ESR1 expression.

All 14,475 genes represented on the bead array were ranked from the most hypermethylated to the most hypomethylated in cluster I with respect to cluster II. The signal-to-noise ratio (the difference in means of the two classes divided by the sum of the standard deviations of the two classes) was used to perform the ranking. When a gene was represented by several probes on the bead array, the most variant one was selected for this analysis. The 20,606 genes represented on the Affymetrix array were ranked according to the same method.

The goal of this GSEA analysis was to determine whether the ESR1 module genes are randomly distributed throughout the ranked lists (suggesting no enrichment of these gene sets in one of the two clusters) or primarily found at the top or bottom (suggesting an enrichment of these gene sets in one of the two clusters). A running sum statistic, corresponding to the enrichment score, was calculated for each gene set on the basis of the ranks of the investigated gene set members, relative to those of the non-members. The significance of such enrichments was estimated by calculating a permutation-based p-value corrected for multiple tests by the false discovery rate (FDR) approach.

This analysis was performed with the freely accessible software GSEA-P, provided by the Broad Institute (http://www.broadinstitute.org/gsea/). This GSEA technique has been described in detail in (Subramanian et al., 2005).

Correlation between methylation and expression data

The correlation between methylation and expression data in the main set of patients was evaluated by Pearson's correlation test between each Infinium methylation probe and the most variant Affymetrix expression probe for the gene concerned. Infinium methylation probes presenting values with a range lower than 20% were excluded from this analysis. The range was calculated by subtracting the smallest methylation value from the greatest one for each probe.

Establishment of the 86 CpG-classifier

To transfer class discovery results from one data set to another in order to independently confirm the results, we used the nearest centroid classification method (Lusa et al., 2007; Sorlie et al., 2003) for assigning new samples of the validation set to one of our 6 clusters. This method is based on the similarity of the DNA methylation profile of a new sample to the DNA methylation profile of the previously identified clusters. A centroid was defined as the vector containing the median methylation values of all the samples assigned to that cluster in the original hierarchical clustering in the main set. For each new sample, a Spearman rank correlation was calculated between its methylation data and the six centroids; the predicted cluster was defined as the category having the highest correlation value. For training the classifier, we excluded those patients in the main set not belonging to any of the 6 most robust clusters. We used the Kruskal-Wallis non parametric test to find the differently methylated CpGs between the six clusters. A ranked CpG list was constructed according to the Kruskal-Wallis test statistic values (see Table

SXI). In order to find the minimal number of CpGs to be used for the nearest centroid classifier, we created different classifiers from this list and calculated the proportion of correctly classified samples from the main set as compared to the original clustering. We started with a classifier using the top 5 CpGs most differentially methylated CpGs between the 6 clusters from this list and added one by one an additional CpG from this list up to a total of 1519 (the number of CpGs for which the FDR-adjusted p-value was 0). At the end, the minimal number of CpGs that yielded the maximum percentage of correct classification (96.38%) was given by 86 (see Figs S7 and S8, and Tables SXII, SXIII and SXIV). Finally, the resulting 86-CpG classifier was applied to the validation dataset to classify the new patients into one of the 6 clusters.

Gene ontology analysis

Gene ontology analysis was done with DAVID (http://david.abcc.ncifcrf.gov/), a web-accessible program providing a comprehensive set of functional annotation tools for understanding the biological meaning of large lists of genes (Huang et al., 2009a). Only genes differentially methylated between each subcluster and normal breast samples and displaying an acceptable anti-correlation between their methylation and expression status (Pearson's coefficient below than -0.4) were selected for this analysis (see also Tables SXX and SXXI). This ensured the selection of genes whose expression is affected by methylation changes, facilitating the biological interpretation of results.

Collection of publicly available gene expression datasets

Gene expression datasets were retrieved from public databases or authors' websites. We used normalized data (log2 intensity in single-channel platforms or log2 ratio in dual-channel platforms). Hybridization probes were mapped to Entrez GeneID as described in (Shi et al., 2006) using RefSeq and Entrez database version 2007.01.21. When multiple

probes were mapped to the same GeneID, the one with the highest variance in a particular dataset was selected. Ten breast cancer microarray datasets were used (Table SXIV). Distant metastasis-free survival (DMFS) was used as survival endpoint. We censored the survival data at 10 years in order to have comparable follow-up across the different studies as described in (Desmedt et al., 2008; Haibe-Kains et al., 2008).

Relapse-free survival analysis

For the meta-analysis performed on publicly available gene expression data, we selected only the genes displaying a high anti-correlation between their methylation and expression status (Pearson's coefficient below than -0.7) in our main set of patients. Among the 85 genes meeting this criterion, several were eliminated because they were not represented on the microarray platforms (9 genes) or because information for these genes was available for less than 700 patients (15 genes). Six other genes were excluded from this meta-analysis because they did not display differential methylation between normal breast samples and IDCs in our population.

The prognostic value of individual CpGs or genes was estimated by univariate Cox regression. Multivariate Cox regression was used to test the independent prognostic values of CpGs or genes of interest in the presence of traditional clinical variables. Cox models were stratified by datasets to account for the possible heterogeneity in patient selection or other potential confounders, as implemented in the 'survival' R package available on CRAN (http://cran.r-project.org/web/packages/survival). The significance of individual hazard ratios was estimated by Wald's test. For univariate analysis, the p-values were corrected for multiple testing by means of the false discovery rate (FDR) and variables with a FDR below than 0.1 were considered prognostic. For multivariate analysis, variables with a p-value below than 0.05 were considered prognostic.

Treatment of breast cancer epithelial cell lines with 5-aza-2'-deoxycytidine

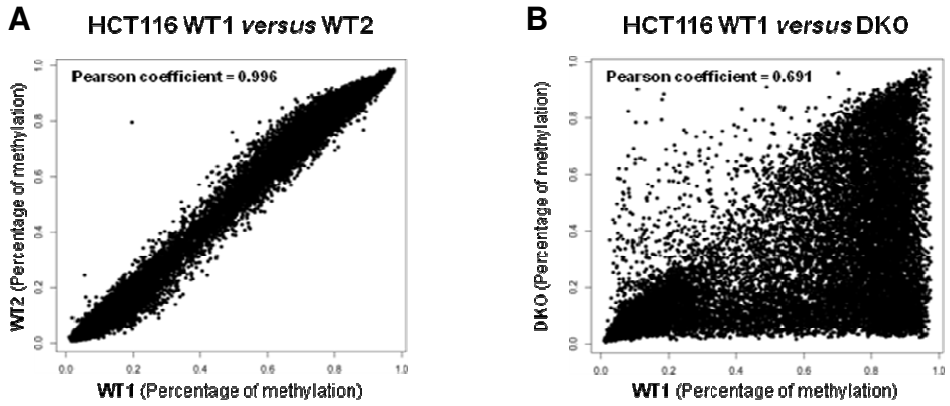
Breast cancer epithelial cell lines MCF-7, MDA-MB-231, MDA-MB-361, T47D, SKBR3, BT20 and ZR-75-1 were treated with 1µM of 5-aza-2'-deoxycytidine (Sigma) during 4 days. Medium containing the drug was refreshed every day.

Additional statistical analyses

Spearman's correlation was used to compare Infinium data with bisulphite genomic sequencing or pyrosequencing data. The Mann-Whitney U test and the Kruskal-Wallis test were used to test for differences of a continuous variable between two or multiple subgroups, respectively. Chi-square tests were used to compare discrete variables and the p-values were estimated by the likelihood ratio or Fisher's Exact test (for comparison of binary variables).

We used the Phi coefficient to determine the strength of associations between the "known expression subtypes" of breast cancer and our DNA methylation-based clusters. The values range from 0 to 1, and can be interpreted in a similar way to Spearman's rank correlation coefficient. The significance of such associations was computed by means of a chi-square test.

Supplemental Figures



C

Gene	Illumina ID	Strand analysed by Infinium	Coding strand	Infinium methylation (%)	Reported methylation data (%; BGS)	Correlation Infinium *vs.* reported methylation data*
SFRP4	cg08261094	Bottom	Bottom	95 ± 1	87 (Ting et al., 2008)	++
SCNN1B	cg23113963	Bottom	Top	93 ± 2	100 (Jacinto et al., 2007)	++
HOXA11	cg15760840	Bottom	Bottom	93 ± 1	94 (Hayashi et al., 2007)	++
HOXA2	cg26069745	Top	Bottom	92 ± 2	88 (Hayashi et al., 2007)	++
DES	cg18182399	Top	Top	90 ± 2	98 (Hayashi et al., 2007)	++
BHLHB9	cg00968475	Top	Top	88 ± 0.3	100 (Jacinto et al., 2007)	++
HOXA6	cg04265576	Bottom	Bottom	87 ± 1	88 (Hayashi et al., 2007)	++
HOXA5	cg12128839	Bottom	Bottom	85 ± 1	95 (Hayashi et al., 2007)	++
BHLHB9	cg15309236	Bottom	Top	83 ± 2	100 (Jacinto et al., 2007)	++
ICAM1	cg08607082	Top	Top	64 ± 1	60 (Ting et al., 2008)	++
ICAM1	cg22874046	Top	Top	57 ± 0.3	30 (Ting et al., 2008)	+
HOXA1	cg12686016	Bottom	Bottom	26 ± 0.5	26 (Hayashi et al., 2007)	++
SNCG	cg21012874	Top	Top	15 ± 1	12 (Ye et al., 2008)	++
HOXA7	cg23432345	Bottom	Bottom	10 ± 3	0 (Hayashi et al., 2007)	++
HOXA4	cg04317399	Top	Bottom	7 ± 2	0 (Hayashi et al., 2007)	++
ATM	cg03165700	Top	Top	7 ± 1	0 (Brandes et al., 2007)	++

* Based on the hypothesis that all reference papers check methylation on the coding strand and that methylation is symmetrical between the two strands.

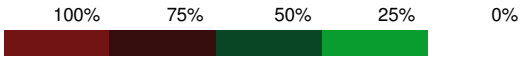


Figure S1, related to Figure 1. Pilot Infinium experiments on HCT116 cells, showing the sensitivity, specificity, and high reproducibility of the technique.

A and B. Scatter plots for two technical replicates of HCT116 WT (A) and for one sample of HCT116 WT *versus* one sample of HCT116 DKO (B). WT: Wild-type cell line; DKO: A double-knockout cell line for the DNMT1 and DNMT3B DNA methyltransferases (Rhee et al., 2002).

C. Methylation status in HCT116 WT of representative CpGs examined by bead array and their correlation with previously reported data. BGS: Bisulphite Genomic Sequencing.

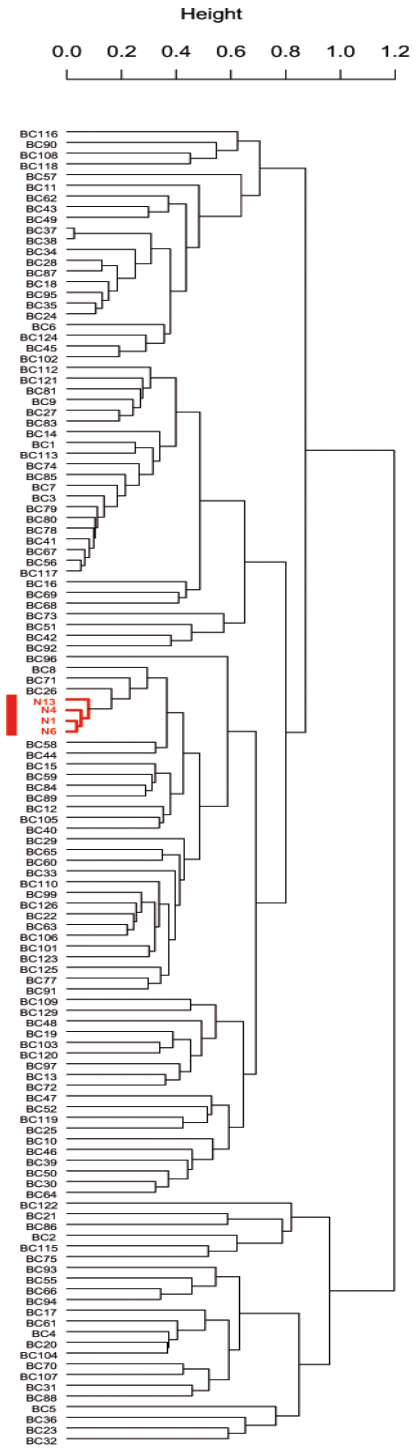


Figure S2, related to Figure 1. Hierarchical analysis of the 123 breast samples of the main patient set showing a grouping of normal samples. Clustering was performed on the 10% most variant CpGs among all samples. BC: Breast Cancer; N: Normal sample (in red).

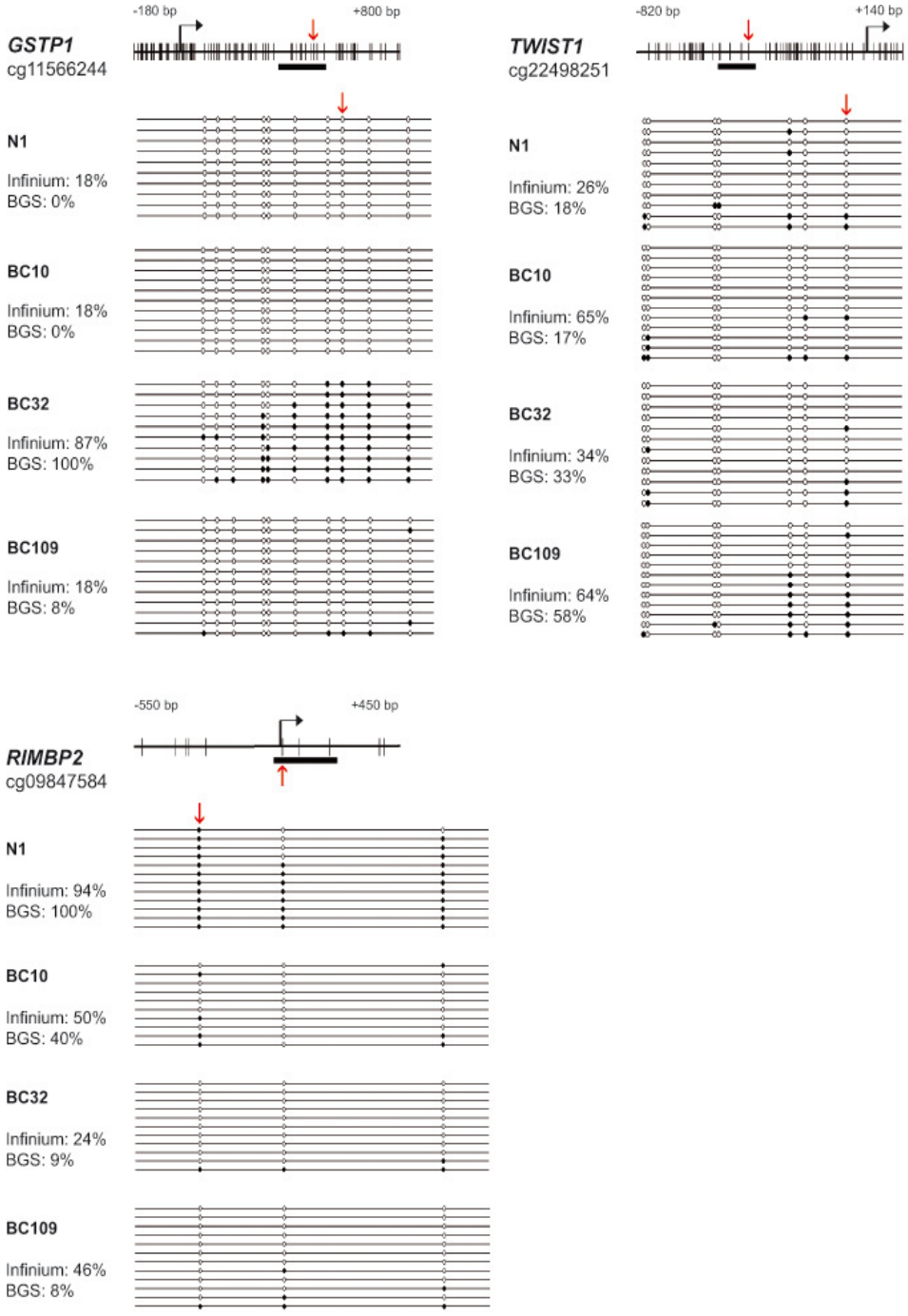


Figure S3, related to Figure 1. Bisulphite genomic sequencing applied to the *GSTP1*, *TWIST1* and *RIMBP2* promoters validating the methylation data obtained by bead array technology. Red arrows indicate the location of the CpGs investigated by means of the bead array. Data are represented as in (Bock et al., 2005). Black and white circles correspond respectively to methylated and unmethylated CpGs. No circle: undetermined sequence.

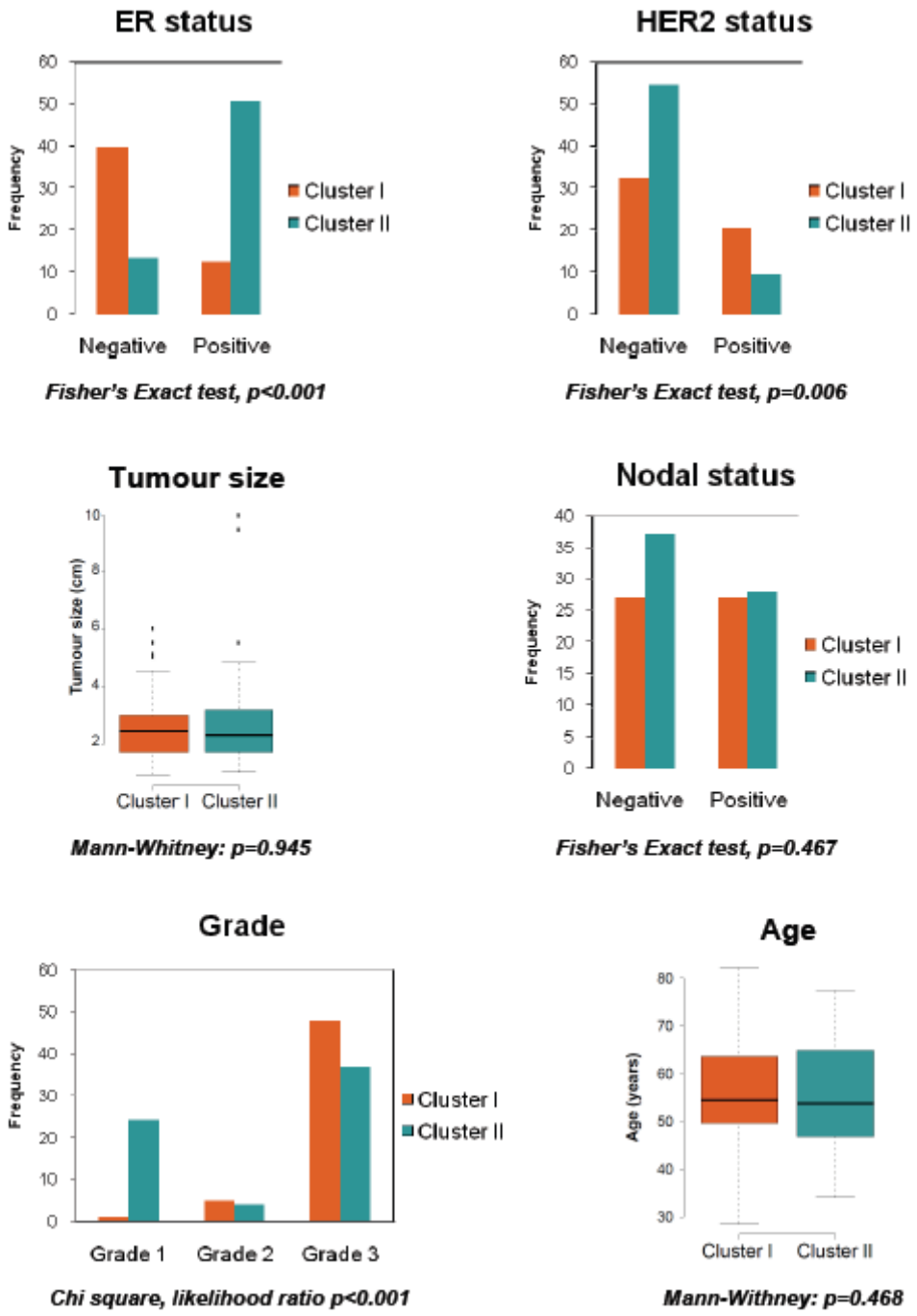


Figure S4, related to Figure 2. Association between methylation clusters I and II of the main patient set and the clinical data. ER-positive tumours were predominant in cluster II, whereas cluster I seemed to contain a moderately higher number of HER2-positive tumours. Grade 1 tumours were grouped in cluster II. No significant association with tumour size, nodal status, or age was found.

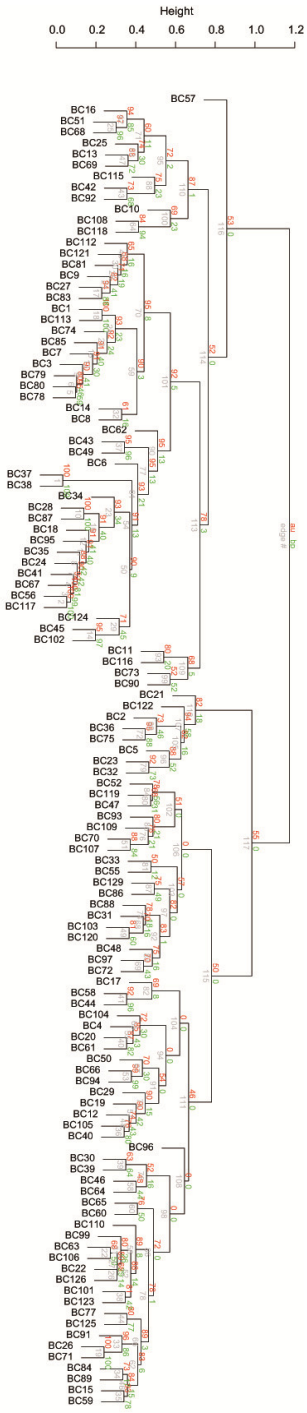


Figure S5, related to Figure 3. Hierarchical clustering of the 119 breast tumours of the main set with all probability stability values. Red values correspond to the probability stability values given by the 'pvclust' package. The 6 methylation clusters selected presented a stability of at least 0.75 and included at least 8 patients (see Supplemental Materials and Methods section for detailed of this analysis).

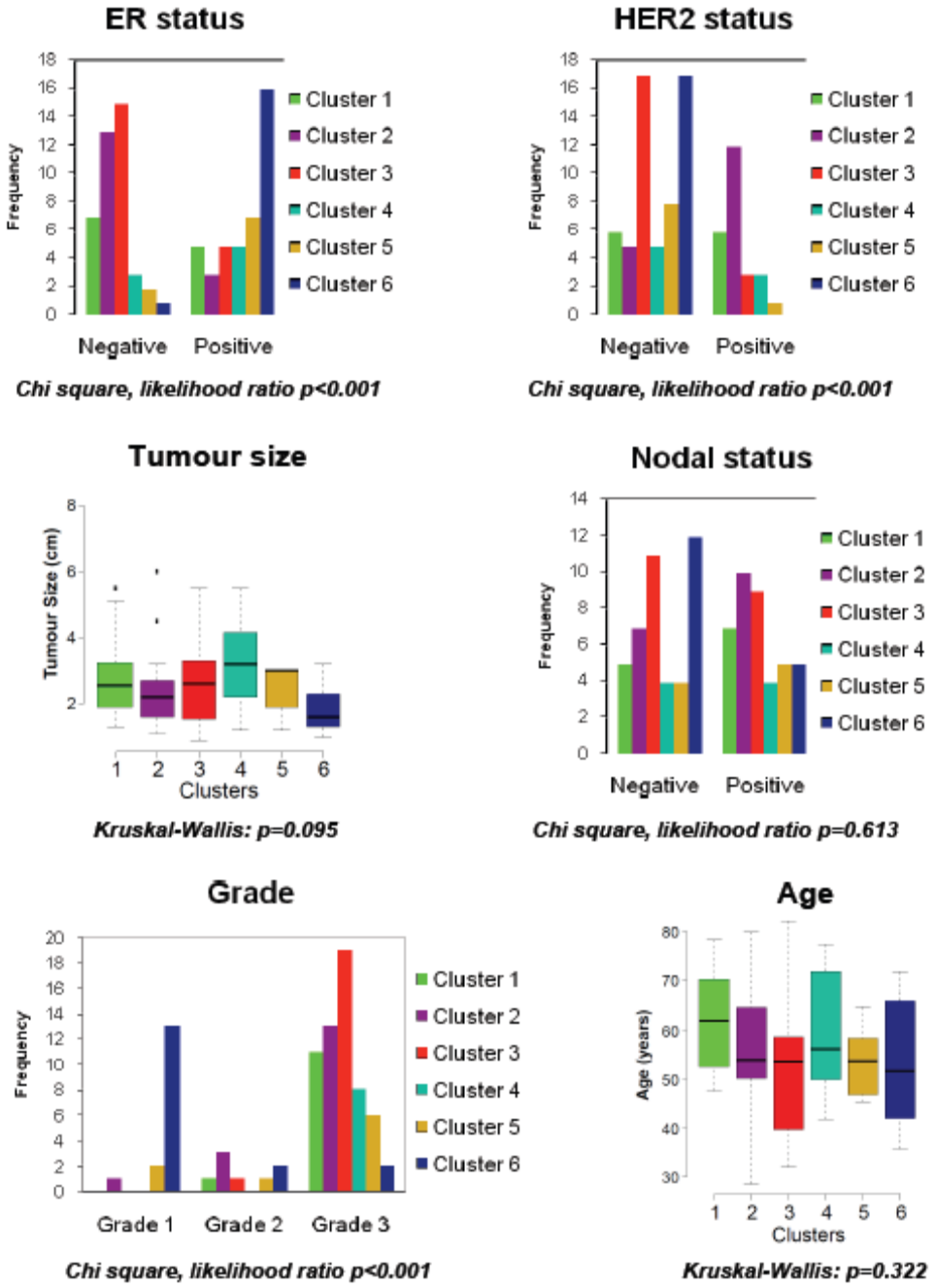


Figure S6, related to Figure 3. Association between methylation clusters 1 to 6 of the main patient set and the clinical data. Cluster 6 contained almost exclusively ER-positive tumours, whereas clusters 2 and 3 were composed principally of ER-negative tumours. HER2-positive tumours were predominant in cluster 2 and HER2-negative tumours were predominant in clusters 3 and 6. Cluster 6 contained almost exclusively grade 1 tumours. No significant association with tumour size, nodal status or age was found.

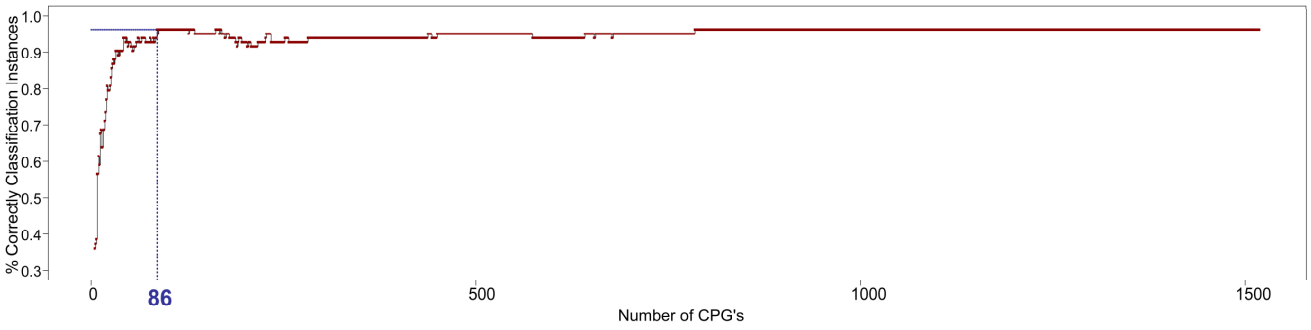


Figure S7, related to Figure 3. Proportion of correctly classified patients in the main set as a function of the number of CpGs included in the classifier. In order to find the minimal number of CpGs to be used for the nearest centroid classifier, we created different classifiers from the list of differentially methylated CpGs between the 6 clusters (see Table SXI) and calculated the proportion of correctly classified samples from the main set as compared to the original clustering. We started with a classifier using the top 5 CpGs most differentially methylated CpGs between the 6 clusters from this list and added one by one an additional CpG from this list up to a total of 1519 (the number of CpGs for which the FDR-adjusted p-value was 0). At the end, the minimal number of CpGs that yielded the maximum percentage of correct classifications (96.38%) was given by 86.

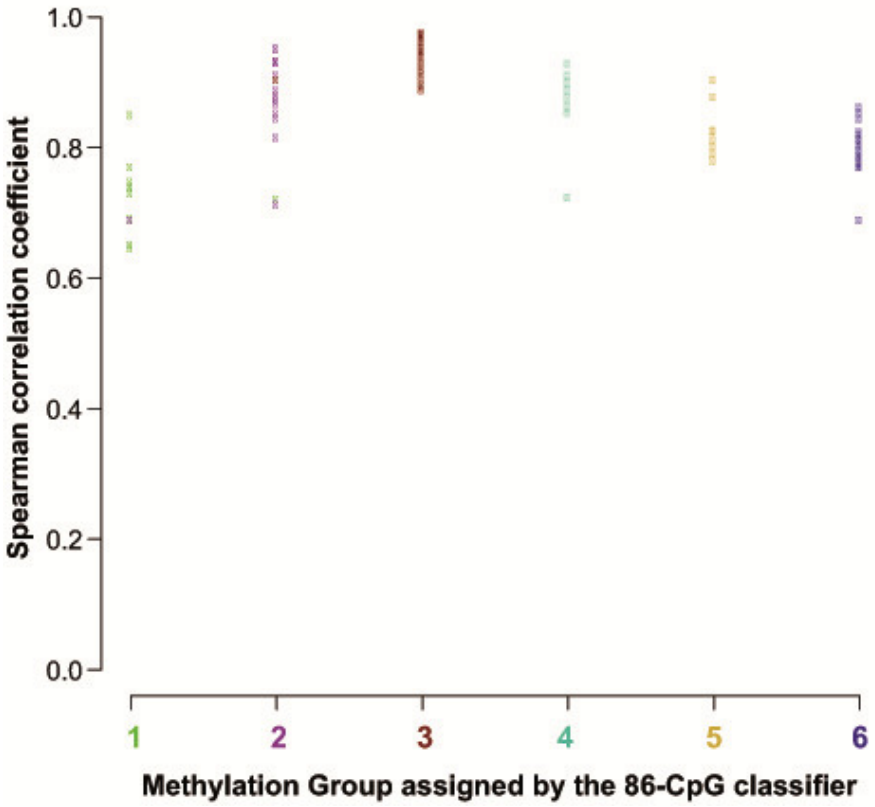


Figure S8, related to Figure 3. Correlation plot of main set tumours with the 6 centroids. Each sample displays the colour of its methylation group assigned by the unsupervised clustering of Fig 3A.

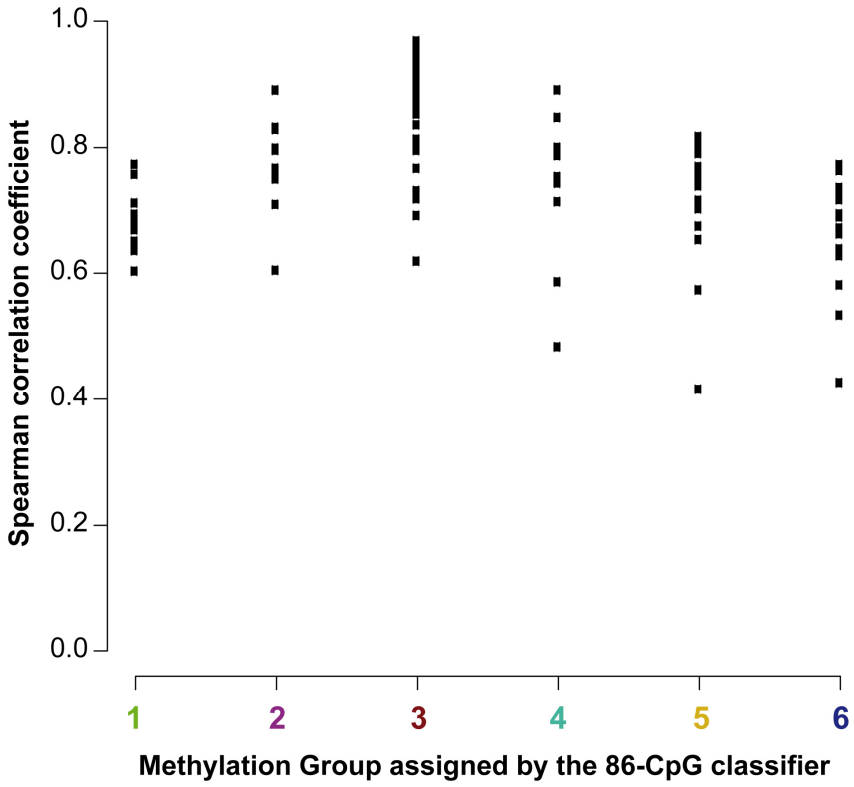


Figure S9, related to Figure 3. Correlation plot of validation set tumours with the 6 centroids. Each sample was placed in the group with which it presented the highest correlation.

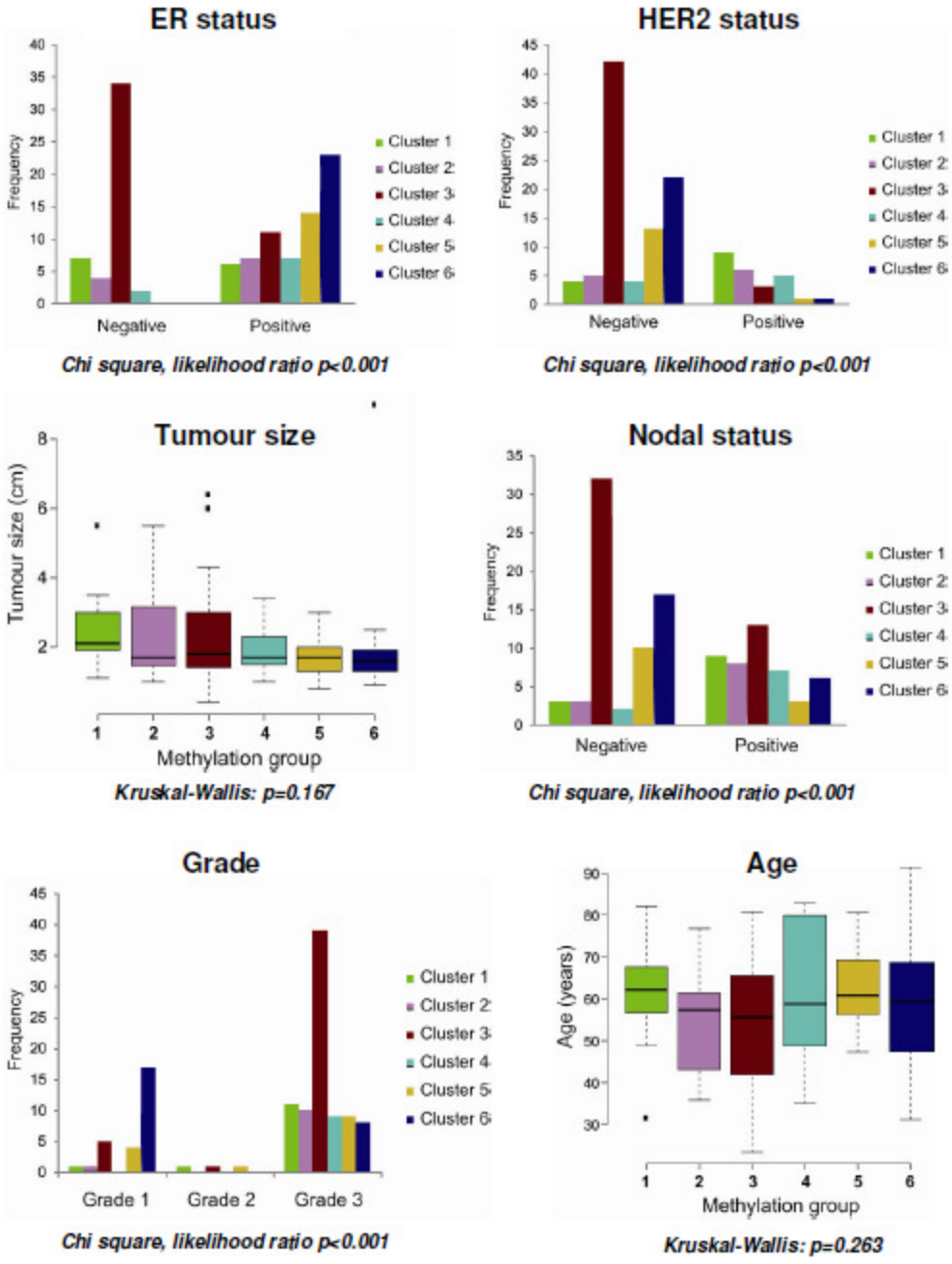


Figure S10, related to Figure 3. Association between the 6 groups of tumours of the validation set and the clinical data. Clusters 5 and 6 contained exclusively ER-positive tumours, whereas clusters 3 were composed principally of ER-negative tumours. HER2-positive tumours were predominant in clusters 1 and 2. Cluster 6 contained majorly grade 1 tumours. No significant association with tumour size or age was found.

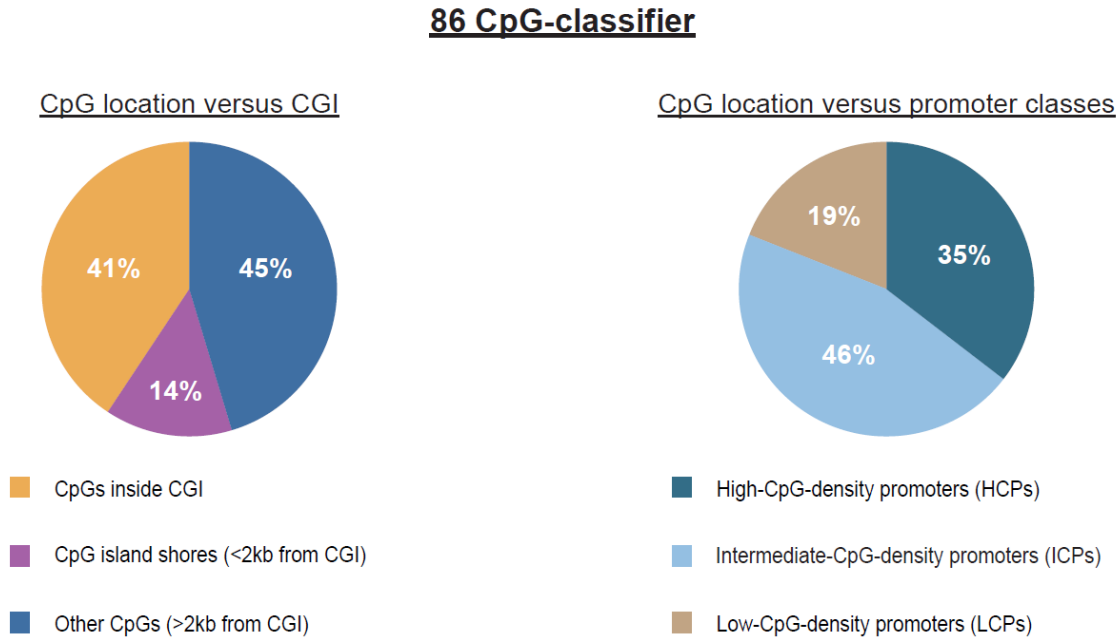


Figure S11, related to Figure 3. Characteristics of the 86 CpG-classifier in terms of CpG location *vs* CGI and *vs* promoter classes.

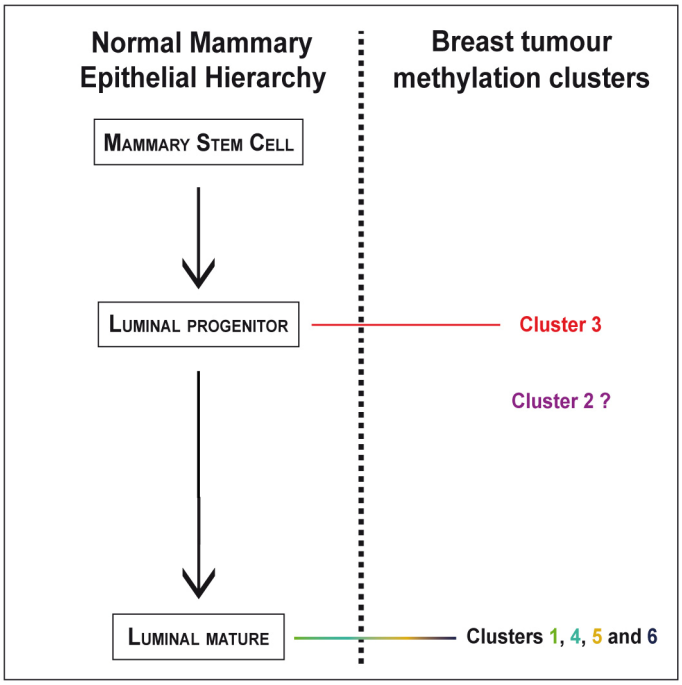


Figure S12, related to Figure 4. Scheme suggesting a different cell type origin for the six methylation clusters identified in the main set of patients. This model derived from the results presented in Figure 4. Cluster 3 tumours showed an expression profile very close to that of luminal progenitor cells, whereas clusters 1, 4, 5, and 6 tumours appeared to be closer to mature luminal cells. These observations suggest that methylation patterns distinguished here might reflect the cell type of origin of the studied tumours.

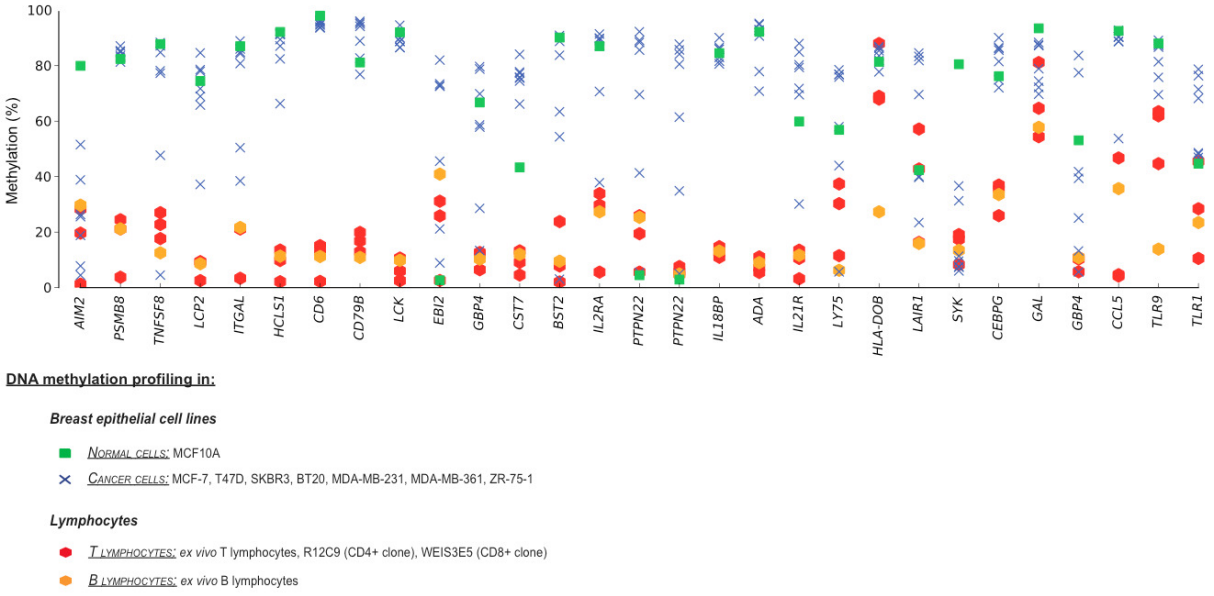


Figure S13, related to Figure 5. Methylation status, as assessed by DNA methylation profiling, of all immune genes revealed by GO analysis in epithelial breast cell lines, *ex vivo* lymphocytes and lymphoid cell lines.

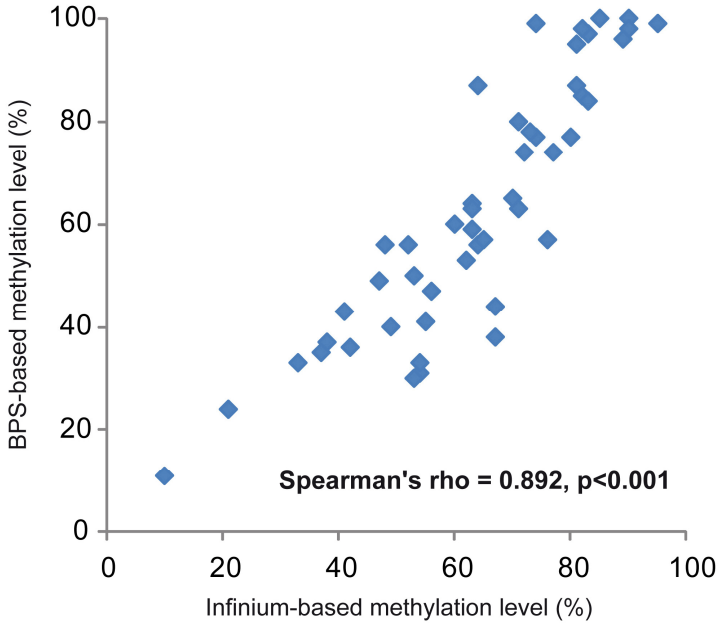


Figure S14, related to Figure 6. Bisulphite pyrosequencing (BPS) of several immune markers highlighted in Figure 6 validating methylation data obtained from Infinium experiment (see also Table SXXXI).

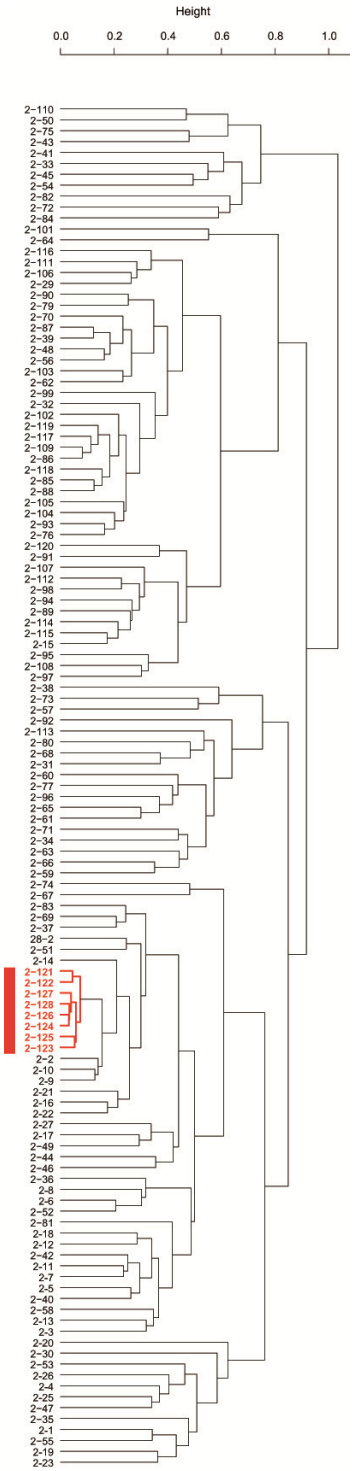


Figure S15. Hierarchical cluster analysis of the 125 breast samples of the validation patient set showing a grouping of normal samples. Clustering was performed on the 10% most variant CpGs among all samples. Normal samples are highlighted in red.

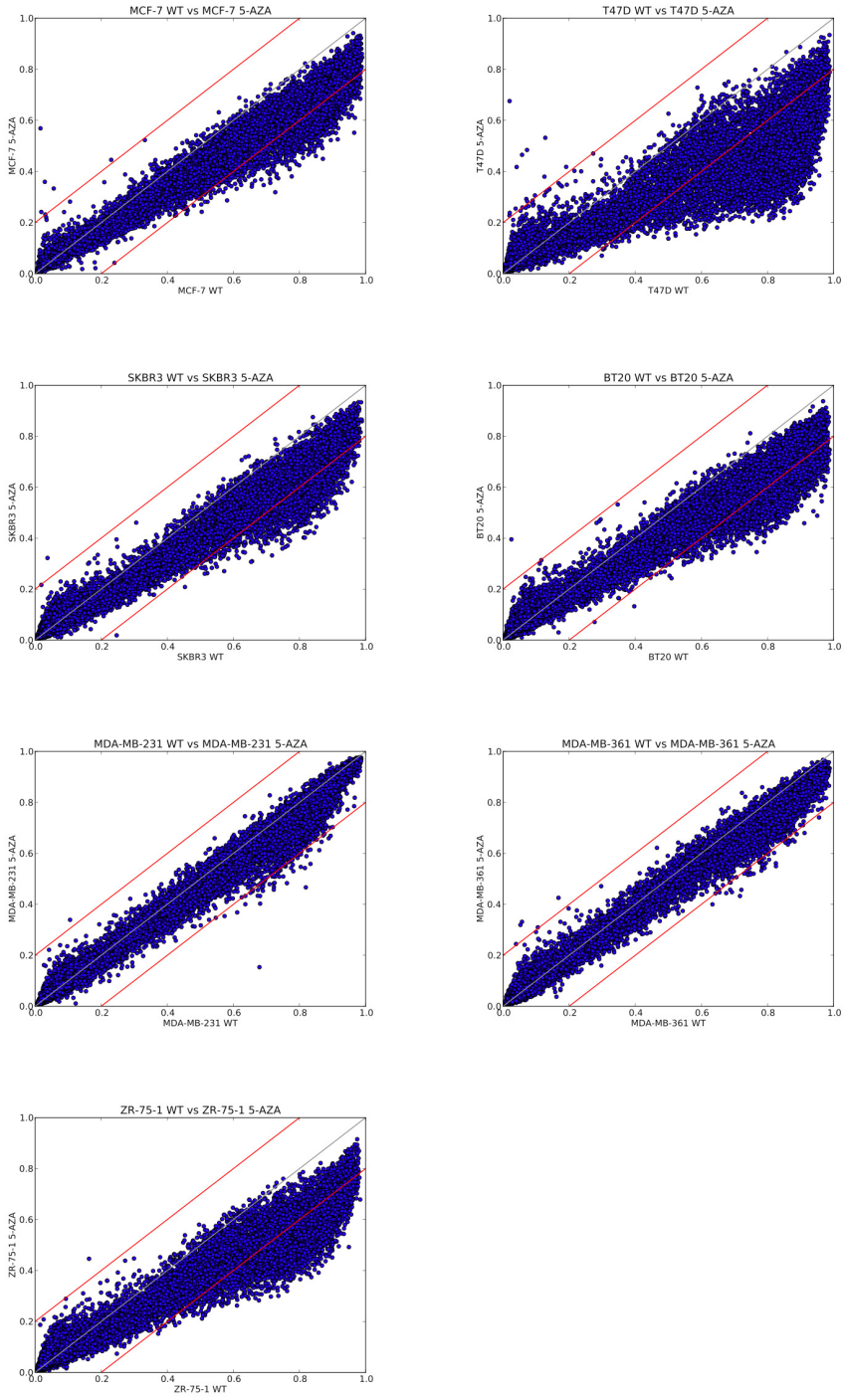


Figure S16. Scatter plots illustrating the overall methylation changes in breast cancer epithelial cell lines treated with 5-aza-2'-deoxycytidine. Red lines indicate a difference of 20% of methylation with a perfect correlation (grey line).

Supplemental Tables

Table SI, related to Figure 1. Characteristics of breast tissue samples of the main patient set.

Characteristic		Number of patients
Tumour size	≤2 cm	44
	>2 cm	75
Nodal status	Negative	64
	Positive	55
Grade	1	25
	2	9
	3	85
ER	Negative	54
	Positive	64
	Unknown	1
HER2	Negative	88
	Positive	31
Subtype IHC	Basal-like	31
	HER2+	31
	Luminal A	25
	Luminal B	32
Subtype GEP	Basal-like	22
	HER2+	21
	Luminal A	23
	Luminal B	22
	Unknown	31
Age	< 50 years	38
	> 50 years	81
Relapse	No	68
	Yes	51

Table SII, related to Figure 1. Demography of breast cancer samples of the main set.

This Table is provided in the additional file **Sup_2.xls**.

Column description:

- Sample_Name: Sample reference.
- GE_QC: 1 and 0 indicate respectively that the sample passed or not the quality control for gene expression profiling. NA indicates that gene expression analysis was not performed on this sample.
- Methyl_QC: 1 indicates that the sample passed the quality control for DNA methylation profiling.
- Subtype_IHC: "Breast cancer expression subtype" determined by IHC as described in the Supplemental Materials and Methods section.
- iTu-ly: Percentage of intratumoral lymphocyte infiltration.
- str-ly: Percentage of stromal lymphocyte infiltration.
- GRADE: Histological grade of the tumour.
- Size_Bin: 1 and 0 indicate, respectively, that the size of the tumour was above or below 2 cm.
- Size_cm: Size of the tumour in cm.
- Nodal_Status: 1 and 0 indicate, respectively, the presence or absence of cancer cells in lymph nodes.
- ER_IHC: ER status determined by IHC. 1 indicates positive; 0 indicates negative.
- HER2_IHC: HER2 status determined by IHC. 1 indicates positive; 0 indicates negative.
- Subtype_GE: "Breast cancer expression subtype" determined by gene expression as described in the Supplemental Materials and Methods section.
- Age_diagnosis: Patient's age (in years) at the time of diagnosis.
- Age_bin: 1 and 0 indicate, respectively, that the patient was above or below 50 years old at the time of diagnosis.
- RFS_event: 1 and 0 indicate, respectively, a relapse event or not.
- RFS_time: Relapse-free survival time in years.
- RFS_event_censored: Relapse-free survival event, censored at 10 years.
- RFS_time_censored: Relapse-free survival time in years, censored at 10 years.
- Relapse_5years: 1 and 0 indicate, respectively, the presence or not of a relapse event within the first 5 years of follow up.
- OS_event: 1 and 0 indicate, respectively, the occurrence or not of an overall survival event.
- OS_time: Overall survival time in years.
- Main_Cluster: Main methylation cluster membership.
- Subcluster: Methylation subcluster membership.

Table SIII, related to Figure 1. Class comparison analysis between IDC and normal breast tissue samples.

This Table is provided in the additional file **Sup_3.xls**. The "All data" tab contains data for all 27,578 CpGs investigated by means of the Infinium bead array. The "HYPER" and "HYPO" tabs are the lists of CpGs that are, respectively, hypermethylated or hypomethylated in IDCs as compared to normal breast tissue samples, according to the criteria described in the Supplemental Materials and Methods section.

Column description:

- Illumina_ID: Illumina probe reference for each investigated CpG.
- SYMBOL: Symbol of the gene concerned.
- Mean.Normal: Mean of the methylation percentage of each CpG for the normal breast samples.
- Median.Invasive: Median of the methylation percentage of each CpG for the IDCs.
- Delta.Beta: Methylation difference between IDCs and normal breast samples for each CpG.
- Proportion.Hyper.20%.Methylation: Percentage of IDCs showing at least 20% hypermethylation as compared to the mean of normal breast samples.
- Proportion.Hypo.20%.Methylation: Percentage of IDCs showing at least 20% hypomethylation as compared to the mean of normal breast samples.
- Wilcox.pVal: p-value given by the Wilcoxon's test.
- Wilcox.pVal.fdr: FDR-corrected Wilcoxon p-value.
- Gene_ID: Gene ID as defined by the NCBI.
- Distance_to_TSS: Distance between the investigated CpG and the transcription start site (in base pairs).
- MapInfo: Position of the investigated CpG on the chromosome.
- CpG_Island_Revisited: 'true', 'shore' and 'false' indicate that the investigated CpG is located inside a CGI, is a CpG island shore, or is neither in a CGI nor a CpG island shore, respectively (according to the definition in (Bock et al., 2007)).
- Promoter_Class: Promoter class based on CpG density and CG content as defined in (Weber et al., 2007). HCP: High-CpG-density promoter; ICP: Intermediate-CpG-density promoter; LCP: Low-CpG-density promoter.

Table SIV, related to Figure 1. Methylation frequencies of representative CpGs provided by this Infinium study and their correlation with previously reported data. MSP: Methylation-Specific PCR ; BPS: Bisulphite PyrosSequencing ; MS-HRM: Methylation-Sensitive High Resolution Melting.

Gene	Illumina ID	Strand analysed by Infinium	Coding strand	Infinium methylation frequency, % (number)[Δ]	Reported methylation data frequency, % (number); technique°	Correlation Infinium *vs.* reported methylation data*
RASSF1A	cg00777121	Top	Bottom	71 (85/119)	70 (19/27); MSP (Fackler et al., 2003)	++
					56 (14/25); MSP (Mehrotra et al., 2004)	++
					58 (52/90); MSP (Feng et al., 2007)	++
	cg08047457	Top	Bottom	72 (86/119)	65 (11/17); MSP (Honorio et al., 2003)	++
	cg21554552	Bottom	Bottom	70 (83/119)	65 (11/17); MSP (Honorio et al., 2003)	++
CCND2	cg25425078	Bottom	Top	9 (11/119)	46 (49/106); MSP (Sharma et al., 2007)	+
					28 (10/36); MSP (Evron et al., 2001)	+
					55 (71/130); MSP (Sunami et al., 2008)	+
APC	cg16970232	Top	Top	39 (46/119)	45 (19/42); MSP (Virmani et al., 2001)	++
					28 (15/54); MSP (Esteller et al., 2000)	++
					39 (51/130);MSP (Sunami et al., 2008)	++
					49 (74/151); MSP (Shinozaki et al., 2005)	++
	cg20311501	Bottom	Top	35 (42/119)	45 (19/42); MSP (Virmani et al., 2001)	++
					28 (15/54); MSP (Esteller et al., 2000)	++
					39 (51/130);MSP (Sunami et al., 2008)	++
					49 (74/151); MSP (Shinozaki et al., 2005)	++
RARβ2	cg27486427	Top	Top	12 (14/119)	17 (15/90); BPS (Feng et al., 2007)	++
					0 (0/21); BPS (Pasquali et al., 2007)	+
	cg26124016	Bottom	Top	4 (5/119)	23 (37/160); MSP (Li et al., 2006)	+
CDH13	cg08747377	Top	Top	17 (20/119)	33 (18/55); MSP (Toyooka et al., 2001)	++
SDHB	cg24305835	Top	Bottom	0 (0/119)	0 (0/72); MS-HRM (Huang et al., 2009b)	++
	cg03861428	Bottom	Bottom	0 (0/119)	0 (0/72); MS-HRM (Huang et al., 2009b)	++
FH	cg06806184	Top	Bottom	0 (0/119)	0 (0/72); MS-HRM (Huang et al., 2009b)	++

[Δ] Each tumour identified as positive shows at least 20% hypermethylation of the indicated CpG site as compared to the mean methylation level of normal samples.

[°] For MSP data, to avoid any discrepancy due to a different location of PCR primers and of the CpG investigated by the Infinium technique, we selected only CpGs included in the primer sequences used for the MSP analyses.

* Based on the hypothesis that all reference papers check methylation on the coding strand and that methylation is symmetrical between the two strands.

Table SV, related to Figure 1. Primers used for bisulphite genomic sequencing.

Gene	PCR round	Sequence 5'-3'	Annealing temperature
CDK3	PCR1	*Forward*: gtttagaggggttttttgattatttg	50°C
		Reverse: aactcctacaactccaaaaaattc	
	PCR2	*Forward*: gagggaatagttggaatgtattttg	45°C
		Reverse: ctaaactactatttcctactaactac	
GSTP1	PCR1	*Forward*: ggtttagagtttttagtatggggtt	50°C
		Reverse: actctaaccctaatctaccaacaa	
	PCR2	*Forward*: aggtaggagtatgtgtttggtag	50°C
		Reverse: tcaaaaatacaaaaaaaaaacaaaa	
TWIST1	PCR1	*Forward*: ggtttggtttttggaattttaaggg	50°C
		Reverse: aaaacaacaatatcattaacctaac	
	PCR2	*Forward*: gtttatttgattattgggtgggttt	50°C
		Reverse: ctataacaacaacaataacaacaac	
RIMBP2	PCR1	*Forward*: aaatatgggggtattattttatatg	50°C
		Reverse: ccttactattaaaaatacaaatacc	
	PCR2	*Forward*: atgaattgaaggatgttatttaggg	50°C
		Reverse: aaacttccaaacaaaaataaccaac	

Table SVI, related to Figure 1. Additional annotation of the Infinium array.

This Table is provided in the additional file **Sup_4.xls** and gives additional information about CpG location.

Column description:

- Illumina_ID: Illumina probe reference for each investigated CpG.
- CpG_Island_UCSC: 'TRUE', 'shore' and 'FALSE' indicate that the investigated CpG is located inside a CGI, is a CpG island shore, or is neither in a CGI nor a CpG island shore, respectively (according to UCSC definition).
- CpG_Island_Revisited: 'true', 'shore' and 'false' indicate that the investigated CpG is located inside a CGI, is a CpG island shore, or is neither in a CGI nor a CpG island shore, respectively (according to the definition in (Bock et al., 2007)).
- Promoter_Class: Promoter class based on CpG density and CG content as defined in (Weber et al., 2007). HCP: High-CpG-density promoter; ICP: Intermediate-CpG-density promoter; LCP: Low-CpG-density promoter.

Table SVII, related to Figure 2. List of CpGs showing a methylation difference of at least 20% between IDC and normal breast samples in at least 30% of IDCs.

This Table, provided in the additional file **Sup_5.xls**, contains all the CpGs used for the clustering presented in Figures 2A and 3A. The percentage of methylation for each of these selected CpGs is given for each of the 119 breast cancer samples.

Column description:

- Illumina_ID: Illumina probe reference for each investigated CpG.
- SYMBOL: Symbol of the gene concerned.
- BCx: Sample reference

Table SVIII, related to Figure 2. List of CpGs differentially methylated between clusters I and II.

This Table is provided in the additional file **Sup_6.xls**. The "All data" tab contains data for all 27,578 CpGs investigated by means of the Infinium bead array. The "I vs II" tab is the list of CpGs differentially methylated between clusters I and II according to the selection criteria described in the Supplemental Materials and Methods section.

Column description:

- Illumina_ID: Illumina probe reference for each investigated CpG.
- SYMBOL: Symbol of the gene concerned.
- Mean.Normal: Mean of the methylation percentage of each CpG for the normal breast samples.
- Median.GR.I: Median of the methylation percentage of each CpG for the main methylation cluster I.
- Median.GR.II: Median of the methylation percentage of each CpG for the main methylation cluster II.
- Delta.Beta: Difference in methylation between the two clusters for each CpG.
- Wilcox.fdr: FDR-corrected Wilcoxon p-value.
- EntrezGene_ID: Gene ID as defined by the NCBI.
- Distance_to_TSS: Distance between the investigated CpG and the transcription start site (in base pairs).
- MapInfo: Position of the investigated CpG on the chromosome.
- CpG_Island_Revisited: 'true', 'shore' and 'false' indicate that the investigated CpG is located inside a CGI, is a CpG island shore, or is neither in a CGI nor a CpG island shore, respectively (according to the definition in (Bock et al., 2007)).
- Promoter_Class: Promoter class based on CpG density and CG content as defined in (Weber et al., 2007). HCP: High-CpG-density promoter; ICP: Intermediate-CpG-density promoter; LCP: Low-CpG-density promoter.

Table SIX, related to Figure 2. GSEA results for the ESR1 module.

This Table is provided in the additional file **Sup_7.xls** and contains two tabs corresponding to the two ESR1 sub-modules, the ESR1 positive and negative modules. Rows in grey indicate genes represented on the Affymetrix expression array but not on the Infinium Methylation bead array.

Column description:

- EntrezGene_ID: Gene ID as defined by the NCBI.
- SYMBOL: Symbol of the gene concerned.
- Affy_ID: Affymetrix probe reference.
- coefficient: Coefficient value indicating the degree of correlation in term of the expression of each gene of this module with ESR1 (see Desmedt et al., 2008).
- Illumina_ID: Illumina probe reference for each investigated CpG.
- Methylation Enrichment: This column indicates whether the gene showed a significant enrichment in cluster I or II in terms of DNA methylation.
- Expression Enrichment: This column indicates whether the gene showed significant enrichment in cluster I or II in terms of expression.
- CpG_Island_Revisited: 'true', 'shore' and 'false' indicate that the investigated CpG is located inside a CGI, is a CpG island shore, or is neither in a CGI nor a CpG island shore, respectively (according to the definition in (Bock et al., 2007)).
- Promoter_Class: Promoter class based on CpG density and CG content as defined in (Weber et al., 2007). HCP: High-CpG-density promoter; ICP: Intermediate-CpG-density promoter; LCP: Low-CpG-density promoter.

Table SX, related to Figure 3. Association between the 6 methylation clusters identified in the main set of patients and the "known expression subtypes". Upper Table indicates the p-values provided by Fisher's Exact test to evaluate the association between each methylation group and each "known expression subtype" determined by immunochemistry (IHC) as well as the Phi value in brackets. Lower Table indicates the likelihood ratio p-values provided by Chi square test to evaluate the association between each methylation group and each "known expression subtype" determined by gene expression (GE) as well as the Phi value in brackets.

		"Known expression subtypes" (IHC)			
		HER2	**Basal-like**	**Luminal A**	**Luminal B**
Methylation groups	**Cluster 1**	0.17 (Phi=0.178)	0.502 (Phi=-0.092)	0.111 (Phi=-0.201)	0.471 (Phi=0.089)
	Cluster 2	<0.001 (Phi=0.448)	1 (Phi=-0.034)	0.172 (Phi=-0.172)	0.009 (Phi=-0.286)
	Cluster 3	0.103 (Phi=-0.186)	<0.001 (Phi=0.491)	0.009 (Phi=-0.275)	0.769 (Phi=-0.054)
	Cluster 4	0.692 (Phi=0.053)	0.675 (Phi=-0.104)	0.344 (Phi=-0.160)	0.091 (Phi=0.198)
	Cluster 5	0.266 (Phi=-0.144)	0.433 (Phi=-0.122)	1 (Phi=0.026)	0.033 (Phi=0.257)
	Cluster 6	0.002 (Phi=-0.333)	0.033 (Phi=-0.237)	<0.001 (Phi=0.736)	0.751 (Phi=-0.077)

		"Known expression subtypes" (GE)			
		HER2	**Basal-like**	**Luminal A**	**Luminal B**
Methylation groups	**Cluster 1**	0.1 (Phi=0.238)	0.059 (Phi=0.250)	0.266 (Phi=0.163)	0.253 (Phi=0.168)
	Cluster 2	<0.001 (Phi=0.445)	0.499 (Phi=0.123)	0.038 (Phi=0.219)	0.327 (Phi=0.149)
	Cluster 3	0.001 (Phi=0.366)	<0.001 (Phi=0.735)	0.004 (Phi=0.315)	0.189 (Phi=0.196)
	Cluster 4	0.592 (Phi=0.113)	0.119 (Phi=0.177)	0.723 (Phi=0.092)	0.477 (Phi=0.134)
	Cluster 5	0.297 (Phi=0.165)	0.027 (Phi=0.256)	0.273 (Phi=0.185)	0.098 (Phi=0.261)
	Cluster 6	0.004 (Phi=0.318)	0.003 (Phi=0.323)	<0.001 (Phi=0.503)	0.087 (Phi=0.254)

Table SXI, related to Figure 3. List of the 2,985 CpGs used for the unsupervised clustering together with their corresponding Kruskal-Wallis test statistics for differential methylation status between clusters 1 to 6.

This Table is provided in the additional file **Sup_8.xls**.

Column description:

- Illumina_ID: Illumina probe reference for each investigated CpG.
- pVal: p-value of the Kruskal-Wallis test for differential methylation between clusters 1 to 6.
- pVal.fdr: FDR-corrected Kruskal-Wallis p-value.

Table SXII, related to Figure 3. Proportion of correctly classified patients as a function of the number of CpGs in the classifier.

This Table is provided in the additional file **Sup_9.xls**.

Column description:

- c.index: concordance index estimate (or percentage of similarity) *i.e.* number of correctly classified patient / total number of patients of main set.
- se: standard error of the estimate.
- upper/lower: upper and lower bound of the confidence interval.
- p.value: p-value of the statistical test (H0: the estimate is different from 0.5).
- No.CpG's: Number of CpG used for the estimation.

Table SXIII, related to Table 1. List of the 86 CpGs of the classifier.

This Table is provided in the additional file **Sup_10.xls**.

Column description:

- Illumina_ID: Illumina probe reference for each investigated CpG.
- SYMBOL: Symbol of the gene concerned.
- CHR: Chromosome concerned.
- MapInfo: Position of the investigated CpG on the chromosome.
- Gene_ID: Gene ID as defined by the NCBI.
- Distance_to_TSS: Distance between the investigated CpG and the transcription start site (in base pairs).
- CpG_Island_Revisited: 'true', 'shore' and 'false' indicate that the investigated CpG is located inside a CGI, is a CpG island shore, or is neither in a CGI nor a CpG island shore, respectively (according to the definition in (Bock et al., 2007)).
- Promoter_Class: Promoter class based on CpG density and CG content as defined in (Weber et al., 2007). HCP: High-CpG-density promoter; ICP: Intermediate-CpG-density promoter; LCP: Low-CpG-density promoter.

Table SXIV, related to Figure 3. Spearman's correlation values for each tumour of the main set with the 6 centroids.

This Table is provided in the additional file **Sup_11.xls**.

Column description:

- Sample_Name: Sample reference.
- Spearman_GrX: Value of the Spearman's correlation coefficient between the indicated sample and the centroid of group X.
- Max_Spearman: Maximum value of the Spearman's coefficient obtained for the indicated sample with one of the 6 centroids.
- Group_Clustering: Methylation group assigned to the indicated sample by the unsupervised clustering.
- Group_Centroid: Methylation group assigned to the indicated sample by the nearest centroid method.

Table SXV, related to Figure 3. Demography of breast cancer samples of the validation set.

This Table is provided in the additional file **Sup_12.xls**.

Column description:

- Sample_Name: Sample reference.
- Methyl_QC: 1 indicates that the sample passed the quality control for DNA methylation profiling.
- Subtype_IHC: "Breast cancer expression subtype" determined by IHC as described in the Supplemental Materials and Methods section.
- iTu-ly: Percentage of intratumoral lymphocyte infiltration.
- str-ly: Percentage of stromal lymphocyte infiltration.
- GRADE: Histological grade of the tumour.
- Size_Bin: 1 and 0 indicate, respectively, that the size of the tumour was above or below 2 cm.
- Size_cm: Size of the tumour in cm.
- Nodal_Status: 1 and 0 indicate, respectively, the presence or absence of cancer cells in lymph nodes.
- ER_IHC: ER status determined by IHC. 1 indicates positive; 0 indicates negative.
- HER2_IHC: HER2 status determined by IHC. 1 indicates positive; 0 indicates negative.
- Age_diagnosis: Patient's age (in years) at the time of diagnosis.
- Age_bin: 1 and 0 indicate, respectively, that the patient was above or below 50 years old at the time of diagnosis.
- RFS_event: 1 and 0 indicate, respectively, a relapse event or not.
- RFS_time: Relapse-free survival time in years.
- Relapse_5years: 1 and 0 indicate, respectively, the presence or not of a relapse event within the first 5 years of follow up.
- OS_event: 1 and 0 indicate, respectively, the occurrence or not of an overall survival event.
- OS_time: Overall survival time in years.
- Methylation_Group: Methylation group assigned to the sample by the 86-CpG classifier.

Table SXVI, related to Figure 3. Spearman's correlation values for each tumour of the validation set with the 6 centroids.

This Table is provided in the additional file **Sup_13.xls**.

Column description:

- Sample_Name: Sample reference.
- Spearman_GrX: Value of the Spearman's correlation coefficient between the indicated sample and the centroid of group X.
- Max_Spearman: Maximum value of the Spearman's coefficient obtained for the indicated sample with one of the 6 centroids.
- Group_Centroid: Methylation group assigned to the indicated sample by the nearest centroid method.

Table SXVII, related to Figure 3. Association between the 6 methylation groups obtained for the validation set of tumours and the "known expression subtypes". The Table indicates the p-values provided by Fisher's Exact test to evaluate the association between each methylation group of the validation set and each "known expression subtype" determined by immunochemistry (IHC) as well as the Phi value in brackets.

		"Known expression subtypes" (IHC)			
		HER2	**Basal-like**	**Luminal A**	**Luminal B**
Methylation groups	**Cluster 1**	<0.001 (Phi=0.413)	0.339 (Phi=-0.112)	0.037 (Phi=-0.194)	0.511 (Phi=-0.083)
	Cluster 2	0.012 (Phi=0.261)	0.170 (Phi=-0.147)	0.453 (Phi=-0.107)	1 (Phi=0.012)
	Cluster 3	0.002 (Phi=-284)	<0.001 (Phi=0.673)	0.023 (Phi=-0.225)	0.017 (Phi=-0.223)
	Cluster 4	0.021 (Phi=0.241)	0.276 (Phi=-0.119)	0.115 (Phi=-0.158)	0.692 (Phi=-0.051)
	Cluster 5	0.296 (Phi=-0.128)	0.01 (Phi=-0.241)	0.735 (Phi=0.048)	0.001 (Phi=0.326)
	Cluster 6	0.014 (Phi=-0.221)	<0.001 (Phi=-0.341)	<0.001 (Phi=0.556)	0.798 (Phi=0.028)

Table SXVIII, related to Figure 5. Lists of CpGs differentially methylated between each of the 6 methylation clusters and normal breast tissue samples in the main set.

This Table is provided in the additional file **Sup_14.xls**. The "All data" tab contains data for all 27,578 CpGs investigated by the Infinium bead array. The 6 "GRx vs N" tabs are lists of CpGs differentially methylated between group x and normal breast samples. The selection criteria used to compile these 6 lists are defined in the Supplemental Materials and Methods section.

Column description:

- Illumina_ID: Illumina probe reference for each investigated CpG.
- SYMBOL: Symbol of the gene concerned.
- Mean.Normal: Mean of the methylation percentage of each CpG for the normal breast samples.
- Median.GRx: Median of the methylation percentage of each CpG for the methylation subcluster x.
- Delta.GRx.vs.N: Methylation difference for each CpG between group x and normal breast samples.
- GRx.pval: p-value given by Wilcoxon's test between group x and the normal group.
- GRx.fdr: FDR-corrected Wilcoxon p-value between group x and the normal group.
- EntrezGene_ID: Gene ID as defined by the NCBI.
- Distance_to_TSS: Distance between the investigated CpG and the transcription start site (in base pairs).
- MapInfo: Position of the investigated CpG on the chromosome.
- CpG_Island_Revisited: 'true', 'shore' and 'false' indicate that the investigated CpG is located inside a CGI, is a CpG island shore, or is neither in a CGI nor a CpG island shore, respectively (according to the definition in (Bock et al., 2007)).
- Promoter_Class: Promoter class based on CpG density and CG content as defined in (Weber et al., 2007). HCP: High-CpG-density promoter; ICP: Intermediate-CpG-density promoter; LCP: Low-CpG-density promoter.

Table SXIX, related to Figure 5. Correlation between DNA methylation and gene expression data in the main set.

This Table is provided in the additional file **Sup_15.xls**.

Column description:

- Illumina_ID: Illumina probe reference for each investigated CpG.
- Affy_ID: Affymetrix probe reference.
- EntrezGene_ID: Gene ID as defined by the NCBI.
- SYMBOL: Symbol of the gene concerned.
- CPG_ISLAND: TRUE indicates that the investigated CpG is located in or close to a CpG island. FALSE indicates that the investigated CpG is not close to a CpG island.
- Pearson_coef: Pearson coefficient of correlation between the methylation status of the indicated CpG and the expression status of the gene concerned determined by taking the most variant Affymetrix probe.
- CpG_Island_Revisited: 'true', 'shore' and 'false' indicate that the investigated CpG is located inside a CGI, is a CpG island shore, or is neither in a CGI nor a CpG island shore, respectively (according to the definition in (Bock et al., 2007)).
- Promoter_Class: Promoter class based on CpG density and CG content as defined in (Weber et al., 2007). HCP: High-CpG-density promoter; ICP: Intermediate-CpG-density promoter; LCP: Low-CpG-density promoter.

Table SXX, related to Figure 5. Lists of genes differentially methylated between each of the 6 methylation clusters and normal samples of the main set that display an anti-correlation between their methylation and expression status.

This Table, provided in the additional file **Sup_16.xls**, gives for each cluster the lists of hypo- and hyper-methylated CpGs and genes (compared to normal samples) displaying an anti-correlation between their methylation and expression status (Pearson's coefficient ≤ -0.4)

Column description:

- GRx_HYPOmethylated: CpGs and associated genes hypomethylated in group x as compared to normal samples and displaying an anti-correlation between their methylation and expression status.
- GRx_HYPERmethylated: CpGs and associated genes hypermethylated in group x as compared to normal samples and displaying an anti-correlation between their methylation and expression status.
- Illumina_ID: Illumina probe reference for each investigated CpG.
- SYMBOL: Symbol of the gene concerned.

Table SXXI, related to Figure 5. Gene Ontology analysis revealing the features of each of the 6 methylation clusters identified for the main set of patients.

This Table is provided in the additional file **Sup_17.xls**. This analysis was performed from the lists given in the Table SXX. Each tab corresponds to one analysis of hypomethylated (HYPO) or hypermethylated (HYPER) genes of the indicated subcluster (GRx).

Column description:

- Category: Original database
- Term: Enriched terms
- Count: Number of genes in the list belonging to the indicated term.
- %: Percentage of genes in the list belonging to the indicated term.
- Genes: Official symbol of the genes concerned.
- List Total: Number of genes in the list tested
- PValue: Modified Fisher Exact P-Value as described by DAVID (http://david.abcc.ncifcrf.gov/).
- FDR: FDR-corrected P-Value.

Table SXXII, related to Figure 5. Spearman correlation between methylation status of immune genes described in Figure 5 and the stromal and intratumoral lymphocyte infiltration.

		intratumoral lymphocyte infiltration		stromal lymphocyte infiltration	
Gene_Name	**Illumina_ID**	**rho**	**p-value**	**rho**	**p-value**
AIM2	cg10636246	-0.378	<0.001	-0.309	0.001
PSMB8	cg16890093	-0.447	<0.001	-0.457	<0.001
TNFSF8	cg27631256	-0.451	<0.001	-0.436	<0.001
LCP2	cg17127769	-0.288	0.003	-0.237	0.014
ITGAL	cg14176836	-0.484	<0.001	-0.452	<0.001
HCLS1	cg00141162	-0.508	<0.001	-0.534	<0.001
CD6	cg09902130	-0.586	<0.001	-0.635	<0.001
CD79B	cg07973967	-0.461	<0.001	-0.468	<0.001
LCK	cg17078393	-0.554	<0.001	-0.584	<0.001
EBI2	cg09626634	-0.243	0.012	-0.377	<0.001
GBP4	cg27285720	-0.379	<0.001	-0.343	<0.001
CST7	cg11804789	-0.436	<0.001	-0.412	<0.001
BST2	cg16363586	-0.163	0.095	-0.144	0.141
IL2RA	cg11733245	-0.324	0.001	-0.287	0.003
PTPN22	cg00916635	-0.391	<0.001	-0.365	<0.001
IL18BP	cg16749930	-0.61	<0.001	-0.626	<0.001
ADA	cg20622019	-0.408	<0.001	-0.33	0.001
IL21R	cg19423311	-0.377	<0.001	-0.173	0.076
LY75	cg10107725	-0.37	<0.001	-0.28	0.004
HLA-DOB	cg04576021	-0.399	<0.001	-0.305	0.001
LAIR1	cg06238491	-0.455	<0.001	-0.317	0.001
SYK	cg23447996	-0.264	0.006	-0.238	0.014
CEBPG	cg15046693	-0.406	<0.001	-0.366	<0.001
GAL	cg04464446	-0.283	0.003	-0.265	0.006
GBP4	cg21365602	-0.503	<0.001	-0.426	<0.001
CCL5	cg10315334	-0.572	<0.001	-0.559	<0.001
TLR9	cg21578541	-0.412	<0.001	-0.395	<0.001
TLR1	cg03430998	-0.567	<0.001	-0.526	<0.001

Table SXXIII, related to Figure 6. Univariate Cox regression analysis on methylation data of the main set.

This Table is provided in the additional file **Sup_18.xls**. This analysis was performed on our methylation data for the 6,309 CpGs differentially methylated between IDC and normal breast tissue samples, described in Table SIII.

Column description:

- SYMBOL: Gene symbol.
- Illumina_ID: Illumina probe reference for each investigated CpG.
- EntrezGene_ID: Gene ID as defined by the NCBI.
- Affy_ID: Affymetrix probe reference.
- hazard.ratio: Hazard ratio as estimated by univariate Cox regression analysis.
- lower and upper: 95% confidence interval for the hazard ratio.
- p.value: Wald test p-value.
- fdr: FDR-corrected Wald test p-value.

Table SXXIV, related to Figure 6. Publicly available gene expression data sets used for the meta-analysis.

The column "Survival" indicates the type of survival data available for each dataset. RFS: Relapse-Free Survival, DMFS: Distant Metastasis-Free Survival, OS: Overall Survival.

Reference	Dataset	Technology	Survival	Patients	Probes
(Minn et al., 2007)	VDX	Affymetrix	RFS, DMFS	344	22,283
(van de Vijver et al., 2002)	NKI	Agilent	RFS, DMFS, OS	345	24,481
(Minn et al., 2005)	MSK	Affymetrix	DMFS	99	22,283
(Sotiriou et al., 2006)	UNT	Affymetrix	RFS, DMFS	137	22,283
(Chin et al., 2006)	CAL	Affymetrix	RFS, DMFS, OS	118	22,283
(Desmedt et al., 2007)	TBG	Affymetrix	RFS, DMFS, OS	198	22,283
(Naderi et al., 2007)	NCH	Agilent	RFS, DMFS, OS	135	17,086
(Schmidt et al., 2008)	MAINZ	Affymetrix	DMFS	200	22,283
(Bos et al., 2009)	EMC2	Affymetrix	DMFS	204	54,675
(Li et al., 2010)	DFHCC	Affymetrix	DMFS	115	54,675

Table SXXV, related to Figure 6. Univariate Cox regression meta-analysis on publicly available gene expression data sets.

This meta-analysis was performed on the genes displaying high anti-correlation between their methylation and expression status (Pearson's coefficient below than -0.7), as described in the Supplemental Materials and Methods. The prognostic value of the classical markers (grade, tumour size, nodal status, age of the patient at diagnosis, ER status) was also evaluated. Lower.95 and Upper.95 indicate the 95% confidence interval of the hazard ratio, and n, the number of patients.

Variable	Hazard.Ratio	lower.95	upper.95	P.value	fdr	n
grade	4.319051475	2.70533636	6.895336906	8.81E-10	0	730
CD37	0.637528005	0.508909569	0.798652612	9.02E-05	0.003	951
LAX1	0.607735237	0.469490691	0.786686777	0.000155589	0.003	755
HCLS1	0.66628668	0.534778159	0.830134762	0.000295162	0.004	951
size	1.775376859	1.283496655	2.455762528	0.00052471	0.005	832
RHOH	0.670647193	0.535050445	0.840607948	0.000527206	0.005	952
CD3G	0.704601714	0.56878791	0.87284481	0.001351572	0.012	952
PTPRCAP	0.693100838	0.549253821	0.874620717	0.002010176	0.015	952
CCR7	0.717640112	0.578403622	0.890394373	0.002571111	0.017	887
ARHGAP25	0.79414017	0.679183693	0.928553814	0.003863567	0.02	950
CCL5	0.733823788	0.594450738	0.905873806	0.003978873	0.02	952
BST2	0.747004293	0.61181789	0.912061288	0.004187743	0.02	945
PSCDBP	0.738332573	0.599602639	0.909160421	0.004279438	0.02	890
CD3D	0.769590125	0.639626249	0.925960999	0.005519609	0.022	952
NME5	0.7465137	0.607158777	0.91785333	0.005553296	0.022	951
HEM1	0.745091977	0.603876135	0.919331005	0.006061245	0.022	951
CENTB1	0.753031335	0.61460319	0.922637891	0.00620265	0.022	952
SLC44A4	0.716555934	0.562123142	0.91341624	0.00711915	0.024	755
ICOS	0.776943611	0.644775259	0.936204307	0.007980999	0.024	950
PPP1R16B	0.757698984	0.616947476	0.930561794	0.008136743	0.024	887
CIDEB	0.765412525	0.618428587	0.947330614	0.01399867	0.04	952
UBASH3A	0.816472324	0.693874277	0.960731761	0.014584306	0.04	952
CD6	0.791045558	0.653436134	0.957634637	0.016220318	0.042	944
TRAF3IP3	0.79027337	0.648137351	0.963579706	0.019981307	0.05	881
DNALI1	0.803318339	0.666106667	0.968794318	0.021922321	0.053	952
PADI3	1.282586832	1.027770903	1.600579446	0.027639763	0.064	950
SIT1	0.786510638	0.632504795	0.978014693	0.030779914	0.064	950
CD52	0.798287393	0.65008143	0.980281442	0.031552946	0.064	949
node	1.854933997	1.051885878	3.271058394	0.032782279	0.064	273

GPR171	0.797959507	0.64844202	0.981952673	0.033006747	0.064	950
MAGEA10	1.251763319	1.018281633	1.538779996	0.033009551	0.064	951
LCK	0.80314799	0.652889033	0.987988251	0.038050335	0.071	951
SP140	0.801792991	0.648901416	0.990708273	0.040712689	0.074	886
CD79B	0.796167392	0.638244197	0.993166126	0.043305166	0.076	951
BIN2	0.814941986	0.664344694	0.999677496	0.049639411	0.085	946
PTPN7	0.792341795	0.626269948	1.002451932	0.05243348	0.087	951
PDZK1	0.813311899	0.654827403	1.010153578	0.061677068	0.1	952
HMGCS2	0.823324053	0.6700983	1.011586651	0.064267705	0.101	946
TRAF1	0.860049164	0.714185188	1.035704152	0.111836932	0.172	952
PIK3CG	0.852864273	0.693732209	1.048498915	0.130918607	0.196	952
CCBP2	0.851353503	0.684907289	1.058249487	0.147091806	0.215	952
CALML5	1.152320561	0.948006825	1.400667843	0.154512732	0.221	946
SCRG1	1.186854771	0.928265972	1.517479138	0.171850684	0.24	952
age	0.843892288	0.634787305	1.121878442	0.242671976	0.331	832
er	0.879914817	0.674422359	1.148019599	0.34581516	0.461	885
S100A1	1.100038426	0.877702372	1.378695761	0.407879927	0.532	887
ACTG2	1.102117932	0.858132785	1.415473174	0.446300424	0.561	952
SCNN1A	0.919786588	0.740823935	1.141981688	0.448825642	0.561	946
CRYAB	1.09273719	0.860375019	1.3878536	0.467187455	0.572	952
LDHC	1.076690314	0.874736682	1.325269714	0.485677672	0.583	950
MIA	0.935507087	0.744206524	1.175982045	0.56789208	0.668	952
SYCP2	1.050297885	0.852423577	1.294105041	0.644966227	0.744	945
KRT20	1.031559368	0.878831436	1.210829161	0.703897252	0.797	951
TNS4	1.030114858	0.842888781	1.258928396	0.771886907	0.852	952
SOX10	0.969305349	0.777727696	1.208074322	0.781407858	0.852	952
CHRNA9	0.973691818	0.790085795	1.199965577	0.802531225	0.855	948
TDRD1	1.033987152	0.784876022	1.362163451	0.812158367	0.855	690
RBP1	0.980931649	0.789362527	1.218992372	0.862125942	0.892	952
TFF1	0.988606991	0.822817223	1.187801805	0.902625469	0.918	942
TFF3	1.010010328	0.830061805	1.228969766	0.92074585	0.921	952

Table SXXVI, related to Figure 6. Spearman correlation between methylation status of immune genes described in Figure 6 and the stromal and intratumoral lymphocyte infiltration.

		intratumoral lymphocyte infiltration		stromal lymphocyte infiltration	
Gene_Name	**Illumina_ID**	**rho**	**p-value**	**rho**	**p-value**
LCK	cg17078393	-0.554	<0.001	-0.584	<0.001
CD3D	cg24841244	-0.480	<0.001	-0.563	<0.001
CD3D	cg07728874	-0.548	<0.001	-0.622	<0.001
CD6	cg07380416	-0.589	<0.001	-0.649	<0.001
CD6	cg09902130	-0.586	<0.001	-0.635	<0.001
ICOS	cg15344028	-0.583	<0.001	-0.579	<0.001
CD3G	cg15880738	-0.480	<0.001	-0.514	<0.001
SIT1	cg15518883	-0.536	<0.001	-0.598	<0.001
BST2	cg16363586	-0.163	0.095	-0.144	0.141
CCL5	cg10315334	-0.572	<0.001	-0.559	<0.001
HCLS1	cg00141162	-0.508	<0.001	-0.534	<0.001
RHOH	cg00804392	-0.123	0.212	-0.262	0.007
RHOH	cg11903057	-0.068	0.489	-0.198	0.041
CD79B	cg07973967	-0.461	<0.001	-0.468	<0.001
UBASH3A	cg00134539	-0.360	<0.001	-0.310	0.001
LAX1	cg10117369	-0.404	<0.001	-0.434	<0.001

Table SXXVII, related to Figure 6. Spearman correlation between expression status of immune genes described in Figure 6 and the stromal and intratumoral lymphocyte infiltration.

		intratumoral lymphocyte infiltration		stromal lymphocyte infiltration	
Gene_Name	**Affy_ID**	**rho**	**p-value**	**rho**	**p-value**
LCK	204891_s_at	0.508	<0.001	0.624	<0.001
CD3D	213539_at	0.472	<0.001	0.606	<0.001
CD6	213958_at	0.451	<0.001	0.582	<0.001
ICOS	210439_at	0.571	<0.001	0.63	<0.001
CD3G	206804_at	0.423	<0.001	0.54	<0.001
SIT1	205484_at	0.545	<0.001	0.642	<0.001
BST2	201641_at	0.033	0.77	0.118	0.297
CCL5	1405_i_at	0.545	<0.001	0.634	<0.001
HCLS1	202957_at	0.471	<0.001	0.542	<0.001
RHOH	204951_at	-0.013	0.907	0.173	0.124
CD79B	205297_s_at	0.563	<0.001	0.613	<0.001
UBASH3A	220418_at	0.434	<0.001	0.551	<0.001
LAX1	207734_at	0.526	<0.001	0.646	<0.001

Table SXXVIII, related to Figure 6. Multivariate Cox regression meta-analysis on publicly available gene expression data sets.

This analysis was performed on the 11 immune genes appearing as good prognostic markers in the univariate Cox regression provided in Table SXXV and displaying a good correlation with stromal and intratumoral infiltration (Tables SXXVI and SXXVII). Lower.95 and Upper.95 indicate the 95% confidence interval of the hazard ratio, and n, the number of patients.

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.782098169	0.57957839	1.055383632	0.107962559	741
size	1.340020576	0.961479484	1.867595902	0.083981212	741
grade	4.398033207	2.686723253	7.199363041	3.85E-09	741
er	0.925961144	0.676930243	1.266606197	0.63032068	741
node	1.993075765	1.136034208	3.496682561	0.016187435	741
SIT1	0.6599917	0.502365102	0.867076638	0.002842138	741

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.947747159	0.666485182	1.347703897	0.765118789	546
size	1.296223628	0.813921483	2.064321596	0.274489122	546
grade	4.923533758	2.464824018	9.834854125	6.32E-06	546
er	0.824491233	0.558241611	1.217726842	0.33207764	546
node	5.23442121	1.237767511	22.13595458	0.024455015	546
LAX1	0.446127817	0.310119717	0.641784505	1.36E-05	546

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.815730376	0.605709362	1.098573158	0.179926027	742
size	1.350261099	0.968961036	1.881608204	0.076108607	742
grade	4.270712254	2.62015025	6.961044754	5.74E-09	742
er	0.898932232	0.655768704	1.232262462	0.507900025	742
node	1.985456613	1.130239988	3.487788438	0.017039196	742
HCLS1	0.602372212	0.460056401	0.788712603	0.000227835	742

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.791016381	0.586069628	1.067632386	0.125464002	743
size	1.336212924	0.957464668	1.864784192	0.088312944	743
grade	4.447305084	2.707212296	7.305863133	3.81E-09	743
er	0.883656243	0.644025948	1.212448594	0.44346137	743
node	2.028490613	1.15797223	3.553430785	0.013408473	743
CD3D	0.667293158	0.543518382	0.819255013	0.000111334	743

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.814972815	0.603243078	1.101016677	0.182534825	741
size	1.455661468	1.04379377	2.030046903	0.026929076	741
grade	4.396887623	2.686037542	7.197449948	3.87E-09	741
er	0.869706949	0.63578294	1.189698764	0.382491166	741
node	1.855844417	1.061416677	3.244869404	0.030079032	741
ICOS	0.640822787	0.520023632	0.789683042	2.97E-05	741

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.843106773	0.623527268	1.140012743	0.267567194	735
size	1.400276591	1.000264809	1.960255439	0.049819954	735
grade	4.103756115	2.4933814	6.754207057	2.79E-08	735
er	0.98494381	0.718402528	1.350377081	0.924928239	735
node	1.96365591	1.107469501	3.481761375	0.020927592	735
CD6	0.875910603	0.739643346	1.037282885	0.124615675	735

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.810235146	0.599268909	1.0954698	0.171489956	742
size	1.350831988	0.967991343	1.885086135	0.076955251	742
grade	4.097163474	2.511916282	6.682845544	1.61E-08	742
er	0.909139677	0.664161613	1.244478657	0.552087671	742
node	2.037337019	1.162122985	3.571689214	0.012972722	742
CD79B	0.664381808	0.502243714	0.878862541	0.004175719	742

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.781222718	0.577860841	1.05615209	0.108527271	742
size	1.355296369	0.971945329	1.889847293	0.073098388	742
grade	4.268909828	2.609544229	6.983438303	7.49E-09	742
er	0.874992826	0.63607609	1.20364915	0.411792841	742
node	1.986145103	1.13538492	3.474392075	0.016173634	742
LCK	0.673584038	0.518662828	0.874779203	0.003044328	742

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.793768255	0.587825226	1.071862885	0.131780585	743
size	1.361230624	0.980008306	1.89074807	0.065840561	743
grade	4.645701264	2.839822777	7.599960255	9.58E-10	743
er	0.777853284	0.561584487	1.077408201	0.130686899	743
node	1.944247797	1.112078104	3.399131305	0.019665701	743
CCL5	0.551404359	0.428004708	0.710381828	4.11E-06	743

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.81183076	0.601704913	1.095336216	0.172537127	743
size	1.353550939	0.969870861	1.889014526	0.07506301	743
grade	4.307262419	2.625996736	7.064940063	7.30E-09	743
er	0.926305947	0.678170929	1.265230741	0.630383585	743
node	1.944462487	1.1116814	3.401095279	0.019747903	743
UBASH3A	0.741503992	0.62442346	0.880537337	0.000647399	743

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	n
age	0.792286599	0.587059106	1.069258699	0.127966947	743
size	1.305194443	0.936821995	1.818416458	0.115431743	743
grade	4.52739965	2.77339849	7.390696887	1.55E-09	743
er	0.833481525	0.606620946	1.145182104	0.261157201	743
node	1.863800138	1.06402145	3.264737712	0.029485291	743
CD3G	0.552580273	0.423133705	0.721627594	1.33E-05	743

Table SXXIX, related to Figure 6. Univariate Cox regression meta-analysis on publicly available gene expression data sets specific for each "known expression subtype".

Lower.95/upper.95, 95% confidence interval of the hazard ratio; n, number of patients.

BASAL-LIKE

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	fdr	n
CD6	0.571415127	0.35980797	0.907470858	0.017721616	0.032784991	213
CCL5	0.601220984	0.379386705	0.952765786	0.030315366	0.053412788	213
CD3G	0.614974481	0.393006583	0.962308592	0.033325393	0.056047253	213
LAX1	0.552834594	0.319001003	0.958072497	0.03463195	0.055712264	178
CD3D	0.599642986	0.363138343	0.99017831	0.045658689	0.070390478	213
age	0.557241661	0.295973189	1.049143235	0.070085346	0.103726313	172
LCK	0.632048217	0.376236164	1.061793059	0.083020423	0.113768734	213
HCLS1	0.694316555	0.449956311	1.071382857	0.099266112	0.131173074	213
grade	2.333835064	0.60915775	8.941503419	0.216206627	0.266654849	155
ICOS	0.765441762	0.47602165	1.230828665	0.270037378	0.322302669	213
er	1.325149161	0.603157506	2.911379334	0.483286797	0.55880034	208
UBASH3A	0.84970099	0.528860792	1.365183019	0.500797496	0.561500251	213
SIT1	0.851938648	0.532926849	1.361911981	0.5031992	0.547599137	213
CD79B	0.864632082	0.524298487	1.425883645	0.568758172	0.601258636	213
node	0.631158808	0.081569127	4.883728148	0.659341077	0.677656114	211
size	0.93955348	0.449321006	1.964654956	0.86842147	0.868421495	172

HER2

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	fdr	n
ICOS	0.665653573	0.520062316	0.85200305	0.001230088	0.002167298	142
node	4.604533941	1.787955465	11.85808776	0.001556726	0.00261813	142
LAX1	0.379778681	0.20236605	0.712727492	0.002575214	0.004142736	105
CD3D	0.517574299	0.306380997	0.87434651	0.013820016	0.020453623	142
LCK	0.533630219	0.318779166	0.893286769	0.01688217	0.024024626	142
CD3G	0.574943427	0.345611487	0.956449529	0.033053232	0.045295168	142
size	1.904053799	1.009143609	3.592571797	0.046804702	0.061849073	126
UBASH3A	0.639066456	0.399576092	1.022098029	0.061659162	0.078668587	142
HCLS1	0.651479447	0.405250274	1.047316924	0.076877637	0.094815753	142
CCL5	0.637778183	0.387309781	1.050221372	0.077159864	0.092094034	142
SIT1	0.656499672	0.410184716	1.050726179	0.079472098	0.091889612	141
CD79B	0.720339802	0.411022928	1.262434273	0.251839036	0.282364994	142
CD6	0.875933541	0.692310708	1.108258994	0.269768688	0.2935718	138
age	1.410285548	0.750438055	2.650325787	0.285499481	0.301813751	126
er	1.106033277	0.63703866	1.920306706	0.720323254	0.740332246	136
grade	1.137095166	0.400598853	3.22763135	0.809271597	0.809271574	106

Luminal A

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	fdr	n
grade	5.162337792	2.065135769	12.90459053	0.000445859	0.000824839	275
size	1.850306583	0.961583288	3.560413844	0.065378974	0.115191519	318
CD3D	0.697135966	0.472866537	1.027771088	0.068507829	0.115217708	345
UBASH3A	0.768113097	0.566321462	1.041807117	0.089776717	0.14442341	345
SIT1	0.663341846	0.408478686	1.077222434	0.09706223	0.14963761	345
CCL5	0.672449535	0.410573335	1.101358365	0.114925908	0.170090348	345
CD79B	0.741453969	0.470759597	1.167801977	0.196817333	0.280086219	344
HCLS1	0.74338516	0.437839466	1.262155511	0.272229064	0.373054653	345
CD3G	0.792669997	0.498933534	1.259337528	0.325256661	0.429803461	345
LAX1	0.753425631	0.414668811	1.368924226	0.352748307	0.450058192	270
CD6	0.871687669	0.520960507	1.458535496	0.601065641	0.741314292	344
LCK	1.080613746	0.681066064	1.714556239	0.742025194	0.857966661	344
er	1.123321638	0.342705919	3.682024241	0.847750681	0.950508296	319
age	0.968467546	0.541901248	1.730812379	0.913873178	0.994509041	318
node	1.046039154	0.288465738	3.793164203	0.945400879	0.999423802	344
ICOS	0.993065905	0.572015048	1.724045364	0.98027602	1.007505894	344

Luminal B

Variable	Hazard.Ratio	Lower.95	Upper.95	P.value	fdr	n
LAX1	0.44407418	0.283660793	0.695203153	0.000385645	0.000713443	209
CD3G	0.529767867	0.354645182	0.791365587	0.001917346	0.003378181	255
HCLS1	0.565073005	0.387754045	0.823479484	0.002970425	0.004995715	254
CD3D	0.609672758	0.432610365	0.85920473	0.00470061	0.007561851	255
LCK	0.603241335	0.420086816	0.866249772	0.006187718	0.009539398	255
UBASH3A	0.553322892	0.350383338	0.873803601	0.011128892	0.01647076	255
CCL5	0.626047812	0.430208929	0.911036093	0.014415646	0.020514574	255
grade	2.774788889	1.191228926	6.463454012	0.018002961	0.024670724	210
SIT1	0.617616772	0.411098071	0.927881943	0.020320012	0.025925532	254
ICOS	0.666539915	0.46455092	0.956354706	0.027648847	0.034100246	255
CD6	0.757102121	0.544668538	1.052389814	0.097710234	0.116621897	255
CD79B	0.764181861	0.529362845	1.10316378	0.151056463	0.174659044	255
size	1.475566638	0.834659682	2.608604382	0.180809598	0.196763396	233
age	0.777738033	0.503583487	1.201144327	0.257001758	0.271687567	233
er	1.524385366	0.6055743	3.837267771	0.370748167	0.381046712	239
node	1.321194737	0.438253574	3.982980711	0.620797266	0.620797276	255

Table SXXX, related to Figure S14. Primers used for bisulphite pyrosequencing.

primer name	primer sequence (5' to 3')
CD3D_EF	TGTGTAAATGTGGTTGTATTGTTAATAGG
CD3D_ER	CATCATATTACTCAAACTAATCTCAAACTCC
CD3D-F2	GTGATTTGGTTTTATTTATTGGATGAGT
CD3D-R2Bio	[Btn]AATAAACCTCACTCCCATCAAT
CD3D-S2	GGTTTTATTTATTGGATGAGTTT
CD3D-S2A-cg077	GGTTTGGTATTGGTTATTTTTT
CD3G_EF	GGTATTTGTATTTGTAGTTTTGTTGAGG
CD3G_ER	TTCTCCTCCATAAAACACTATTTCTCTC
CD3G-F1	TGATGGGTGGAGTTAGTTTAGT
CD3G-R1Bio	[Btn]AAACCCTTCCCCTATTCCATA
CD3G-S1	GGTTGGTTGTTAAGGG
CD6_EF2	GGGGAAGTGTGTTTGTATGGATG
CD6_ER	AAACCACATATCTAAAACTATCTCTAACTACTAC
CD6-F1	AGGTAGTTGGGGTTTTTTTTATTAG
CD6-R1Bio	[Btn]CTACCCTTTACTATTCTTATTCCTATATC
CD6-S1	ATATTTATAGGTTGGGTTTG
CD79B_EF	TAGGTAGGAGAGGAATTGGGGTTATAG
CD79B_ER	CATCCACAAAAAACCCCAACTATACTAC
CD79B-F1	AGTTGGAGATGAGAGTAAATTTTATAGG
CD79B-R1Bio	[Btn]AATACCTCCCCTAAATCCCAATTTACAT
CD79B-S1	GGTTGGGTATAGGAGATA
HCLS1_EF	TTATTGTTAAAATTTTGTAAAAGATTAGGTATAG
HCLS1_ER	TTCCTCCTCAACTCTTACTCTATATTTCC
HCLS1-F1	AGGATGGGGTGGTAGGAAAT
HCLS1-R1Bio	[Btn]CCTCCACCTATACAAACCTCTATTCTA
HCLS1-S1	GGGTGGTAGGAAATG
ICOS_EF	TAAGTAGGTAATTTAAAAATTTAATGGTTTGATG
ICOS_ER	CCTCTATCTTCAAAATCATCAATAATCCATAC
ICOS-F1	GAGGTTTGATTTTATGTTTGTTAGAAATAG
ICOS-R1Bio	[Btn]TCCCAAAAAACCCACTTCC
ICOS-S1	TTTGTTAGAAATAGTTAATAGTTTT
LCK_EF	GGTTTATGGTGGTAGGAAGTTTGG
LCK_ER	TTAACACCTAACTATCCATATACCTAATATCC
LCK-F1	GTTAGGTTAGGTTAGGAGGATTAT
LCK-R1Bio	[Btn]CCAACCACAAAAAACTACTACATC
LCK-S2	GAGAGTTGGTATTGGGGG
SIT1_EF	GTAGTGTGTTTGTGGATTTTTATATTTGTAG
SIT1_ER	ATCTAATCAACAACTTATCCTTCCTCCTAC
SIT1-F1	GTGGGTTTTTTTAGGGGTTGTGA
SIT1-R1Bio	[Btn]TCTCAATCAACCCATCCCTATTA
SIT1-S1	GTTGTGAAGTTGTTATTTTTTATTT

UBASH3A-EF2	TGGTGGAAATAGTTAGGATTGGTG
UBASH3A-ER	CAATATCTTACCCTACAAAATACACTACTTTAAC
UBASH3A-F1	GGTTTAAGGGTAGGAAGAGATGG
UBASH3A-R1Bio	[Btn]ACTAACTAAACCCCCAAATCTCTAAACAAT
UBASH3A-S1	GTAGGAAGAGATGGTAG

Table SXXXI, related to Figure 6. Validation by BPS of methylation values obtained by Infinium experiment for several immune genes highlighted in Figure 6.

Gene Name	Sample	Methylation value by Infinium (%)	Methylation value by BPS (%)
LCK	BC97	83	97
LCK	BC24	54	31
LCK	N4	80	77
LCK	BC85	42	36
LCK	BC56	41	43
CD3G	BC66	74	77
CD3G	BC92	73	78
CD3G	N6	72	74
CD3G	BC3	55	41
CD3G	BC42	71	80
CD6 cg07380416	N13	85	100
CD6 cg07380416	BC122	90	100
CD6 cg07380416	BC24	67	44
CD6 cg07380416	BC80	54	33
CD6 cg09902130	N13	90	98
CD6 cg09902130	BC122	95	99
CD6 cg09902130	BC24	67	38
CD6 cg09902130	BC80	53	30
ICOS	BC79	53	50
ICOS	BC122	82	98
ICOS	BC3	48	56
ICOS	N4	81	87
HCLS1	N1	52	56
HCLS1	BC99	60	60
HCLS1	BC31	63	64
HCLS1	BC7	10	11
HCLS1	BC35	21	24
SIT1	BC2	89	96
SIT1	BC66	81	95
SIT1	N4	82	85
SIT1	BC1	64	87
UBASH3A	BC92	63	59
UBASH3A	BC85	49	40
UBASH3A	N4	64	56
UBASH3A	BC42	63	63
CD79B	N1	74	99

CD79B	BC66	76	57
CD79B	BC7	38	37
CD79B	BC27	56	47
CD79B	BC99	70	65
ITGAL	BC31	65	57
ITGAL	BC18	33	33
ITGAL	BC27	37	35
ITGAL	BC125	71	63
ITGAL	N6	62	53
CD3D	BC92	77	74
CD3D	BC85	47	49
CD3D	N4	83	84

Supplemental references

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a Practical and powerful approach to multiple testing. J R Stat Soc *57*, 289-300.

Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K. L. (2009). Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics *1*, 177-200.

Bock, C., Reither, S., Mikeska, T., Paulsen, M., Walter, J., and Lengauer, T. (2005). BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. Bioinformatics *21*, 4067-4068.

Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG island mapping by epigenome prediction. PLoS Comput Biol *3*, e110.

Bos, P. D., Zhang, X. H., Nadal, C., Shu, W., Gomis, R. R., Nguyen, D. X., Minn, A. J., van de Vijver, M. J., Gerald, W. L., Foekens, J. A., and Massague, J. (2009). Genes that mediate breast cancer metastasis to the brain. Nature *459*, 1005-1009.

Brandes, J. C., Carraway, H., and Herman, J. G. (2007). Optimal primer design using the novel primer design program: MSPprimer provides accurate methylation analysis of the ATM promoter. Oncogene *26*, 6229-6237.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T.*, et al.* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell *10*, 529-541.

Denkert, C., Loibl, S., Noske, A., Roller, M., Muller, B. M., Komor, M., Budczies, J., Darb-Esfahani, S., Kronenwett, R., Hanusch, C.*, et al.* (2010). Tumour-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol *28*, 105-113.

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res *14*, 5158-5165.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S.*, et al.* (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res *13*, 3207-3214.

Esteller, M., Sparks, A., Toyota, M., Sanchez-Cespedes, M., Capella, G., Peinado, M. A., Gonzalez, S., Tarafa, G., Sidransky, D., Meltzer, S. J.*, et al.* (2000). Analysis of adenomatous polyposis coli promoter hypermethylation in human cancer. Cancer Res *60*, 4366-4371.

Evron, E., Umbricht, C. B., Korz, D., Raman, V., Loeb, D. M., Niranjan, B., Buluwela, L., Weitzman, S. A., Marks, J., and Sukumar, S. (2001). Loss of cyclin D2 expression in the majority of breast cancers is associated with promoter hypermethylation. Cancer Res *61*, 2782-2787.

Fackler, M. J., McVeigh, M., Evron, E., Garrett, E., Mehrotra, J., Polyak, K., Sukumar, S., and Argani, P. (2003). DNA methylation of RASSF1A, HIN-1, RAR-beta, Cyclin D2 and Twist in in situ and invasive lobular breast carcinoma. Int J Cancer *107*, 970-975.

Feng, W., Shen, L., Wen, S., Rosen, D. G., Jelinek, J., Hu, X., Huan, S., Huang, M., Liu, J., Sahin, A. A.*, et al.* (2007). Correlation between CpG methylation profiles and hormone receptor status in breast cancers. Breast Cancer Res *9*, R57.

Haibe-Kains, B., Desmedt, C., Sotiriou, C., and Bontempi, G. (2008). A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics *24*, 2200-2208.

Hayashi, H., Nagae, G., Tsutsumi, S., Kaneshiro, K., Kozaki, T., Kaneda, A., Sugisaki, H., and Aburatani, H. (2007). High-resolution mapping of DNA methylation in human genome using oligonucleotide tiling array. Hum Genet *120*, 701-711.

Honorio, S., Agathanggelou, A., Schuermann, M., Pankow, W., Viacava, P., Maher, E. R., and Latif, F. (2003). Detection of RASSF1A aberrant promoter hypermethylation in sputum from chronic smokers and ductal carcinoma in situ from breast cancer patients. Oncogene *22*, 147-150.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Huang, K. T., Dobrovic, A., and Fox, S. B. (2009b). No evidence for promoter region methylation of the succinate dehydrogenase and fumarate hydratase tumour suppressor genes in breast cancer. BMC Res Notes *2*, 194.

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M.*, et al.* (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet *41*, 178-186.

Jacinto, F. V., Ballestar, E., Ropero, S., and Esteller, M. (2007). Discovery of epigenetically silenced genes by methylated DNA immunoprecipitation in colon cancer cells. Cancer Res *67*, 11481-11486.

Li, S., Rong, M., and Iacopetta, B. (2006). DNA hypermethylation in breast cancer and its association with clinicopathological features. Cancer Lett *237*, 272-280.

Li, Y., Zou, L., Li, Q., Haibe-Kains, B., Tian, R., Desmedt, C., Sotiriou, C., Szallasi, Z., Iglehart, J. D., Richardson, A. L., and Wang, Z. C. (2010). Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. Nat Med *16*, 214-218.

Lusa, L., McShane, L. M., Reid, J. F., De Cecco, L., Ambrogi, F., Biganzoli, E., Gariboldi, M., and Pierotti, M. A. (2007). Challenges in projecting clustering results across gene expression-profiling datasets. J Natl Cancer Inst *99*, 1715-1723.

Mehrotra, J., Vali, M., McVeigh, M., Kominsky, S. L., Fackler, M. J., Lahti-Domenici, J., Polyak, K., Sacchi, N., Garrett-Mayer, E., Argani, P., and Sukumar, S. (2004). Very high frequency of hypermethylated genes in breast cancer metastasis to the bone, brain, and lung. Clin Cancer Res *10*, 3104-3109.

Minn, A. J., Gupta, G. P., Padua, D., Bos, P., Nguyen, D. X., Nuyten, D., Kreike, B., Zhang, Y., Wang, Y., Ishwaran, H.*, et al.* (2007). Lung metastasis genes couple breast tumour size and metastatic spread. Proc Natl Acad Sci U S A *104*, 6740-6745.

Minn, A. J., Gupta, G. P., Siegel, P. M., Bos, P. D., Shu, W., Giri, D. D., Viale, A., Olshen, A. B., Gerald, W. L., and Massague, J. (2005). Genes that mediate breast cancer metastasis to lung. Nature *436*, 518-524.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E.*, et al.* (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet *34*, 267-273.

Naderi, A., Teschendorff, A. E., Barbosa-Morais, N. L., Pinder, S. E., Green, A. R., Powe, D. G., Robertson, J. F., Aparicio, S., Ellis, I. O., Brenton, J. D., and Caldas, C. (2007). A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene *26*, 1507-1516.

Pasquali, L., Bedeir, A., Ringquist, S., Styche, A., Bhargava, R., and Trucco, G. (2007). Quantification of CpG island methylation in progressive breast lesions from normal to invasive carcinoma. Cancer Lett *257*, 136-144.

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J.*, et al.* (2010). The UCSC Genome Browser database: update 2010. Nucleic Acids Res *38*, D613-619.

Rhee, I., Bachman, K. E., Park, B. H., Jair, K. W., Yen, R. W., Schuebel, K. E., Cui, H., Feinberg, A. P., Lengauer, C., Kinzler, K. W.*, et al.* (2002). DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. Nature *416*, 552-556.

Schmidt, M., Bohm, D., von Torne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H. A., Hengstler, J. G., Kolbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res *68*, 5405-5413.

Sharma, G., Mirza, S., Prasad, C. P., Srivastava, A., Gupta, S. D., and Ralhan, R. (2007). Promoter hypermethylation of p16INK4A, p14ARF, CyclinD2 and Slit2 in serum and tumour DNA from breast cancer patients. Life Sci *80*, 1873-1881.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., *et al.* (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol *24*, 1151-1161.

Shinozaki, M., Hoon, D. S., Giuliano, A. E., Hansen, N. M., Wang, H. J., Turner, R., and Taback, B. (2005). Distinct hypermethylation profile of primary breast cancer is associated with sentinel lymph node metastasis. Clin Cancer Res *11*, 2156-2162.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., *et al.* (2003). Repeated observation of breast tumour subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A *100*, 8418-8423.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., *et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst *98*, 262-272.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Sunami, E., Shinozaki, M., Sim, M. S., Nguyen, S. L., Vu, A. T., Giuliano, A. E., and Hoon, D. S. (2008). Estrogen receptor and HER2/neu status affect epigenetic differences of tumour-related genes in primary breast tumours. Breast Cancer Res *10*, R46.

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics *22*, 1540-1542.

Ting, A. H., Suzuki, H., Cope, L., Schuebel, K. E., Lee, B. H., Toyota, M., Imai, K., Shinomura, Y., Tokino, T., and Baylin, S. B. (2008). A requirement for DICER to maintain full promoter CpG island hypermethylation in human cancer cells. Cancer Res *68*, 2570-2575.

Toyooka, K. O., Toyooka, S., Virmani, A. K., Sathyanarayana, U. G., Euhus, D. M., Gilcrease, M., Minna, J. D., and Gazdar, A. F. (2001). Loss of expression and aberrant methylation of the CDH13 (H-cadherin) gene in breast and lung carcinomas. Cancer Res *61*, 4556-4560.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., *et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med *347*, 1999-2009.

Virmani, A. K., Rathi, A., Sathyanarayana, U. G., Padar, A., Huang, C. X., Cunnigham, H. T., Farinas, A. J., Milchgrub, S., Euhus, D. M., Gilcrease, M., *et al.* (2001). Aberrant methylation of the adenomatous polyposis coli (APC) gene promoter 1A in breast and lung carcinomas. Clin Cancer Res *7*, 1998-2004.

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet *39*, 457-466.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schutz, F.*, et al.* (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res *10*, R65.

Ye, Q., Zheng, M. H., Cai, Q., Feng, B., Chen, X. H., Yu, B. Q., Gao, Y. B., Ji, J., Lu, A. G., Li, J. W.*, et al.* (2008). Aberrant expression and demethylation of gamma-synuclein in colorectal cancer, correlated with progression of the disease. Cancer Sci *99*, 1924-1932.